STRUTTURE, SIMBOLI E MENTI

Ontologia del Significato tra Matematica, Linguaggio, Cognizione e Sistemi Artificiali

Avvertenza metodologica

Nella storia della filosofia ci s'imbatte sovente in un tipo di quesito che resiste ad ogni tentativo di soluzione — non perché le risposte fornite siano state insufficienti, ma perché la domanda stessa muta forma ad ogni risposta che pretende di affrontarla. Il quesito di cui qui ci si occupa appartiene proprio a quella specie di domande: in che senso le strutture formali — le strutture che la matematica descrive, il linguaggio articola, la mente elabora e i sistemi artificiali simulano — sono portatrici di contenuto semantico? Che cosa significa, per una struttura, significare qualcosa?

La difficoltà non è tecnica in senso ristretto. Non si tratta di affinare una notazione, di aggiornare un assioma, di aggiungere un lemma a un sistema già consolidato. Si tratta di qualcosa di più radicale: la struttura del rapporto tra la forma e il contenuto — tra la sintassi che si può vedere e manipolare e la semantica che si pretende di catturare — sfugge ad ogni tentativo di riduzione che parta da un solo dei due lati. Partire dalla struttura formale conduce al problema del grounding: come fanno i simboli astratti a significare qualcosa? Partire dal contenuto semantico porta al problema della codificazione: come si cattura il significato in una struttura finita di segni discreti? Entrambi i cammini si chiudono su se stessi prima di raggiungere il fondo.

Questo scritto non è, dunque, né una rassegna bibliografica né una sintesi divulgativa. Esso si propone di percorrere le quattro regioni concettuali che abitano il problema — l'ontologia matematica, la filosofia del linguaggio, le scienze cognitive, l'intelligenza artificiale — non come territori separati da visitare in sequenza, ma come angolature diverse sulla stessa cosa: il mistero del significato che abita le strutture. Il percorso non è lineare, né converge verso una tesi unica. La fertilità intellettuale di questi domini risiede esattamente nell'inesauribilità delle aporie che sollevano. Le posizioni in conflitto, le dimostrazioni di impossibilità, i progressi tecnici che precisano senza sciogliere — tutto ciò è trattato con eguale serietà, perché eguale è la loro importanza per capire dove ci troviamo.

PARTE PRIMA: L'ONTOLOGIA DELLE STRUTTURE MATEMATICHE

I.1 — Il problema dell'esistenza matematica: genealogia di un'aporia permanente

Consideriamo la proposizione «la radice quadrata di 2 è irrazionale». Essa è vera. Ma in quale senso? Descrive un fatto riguardante un oggetto — l'oggetto $\sqrt{2}$ — che esiste da qualche parte? O è una verità puramente interna a un sistema formale, priva di referente esterno? O è, forse, la descrizione di una costruzione che la mente compie secondo procedure determinate? Queste tre risposte non sono varianti di una stessa posizione, ma costituiscono posizioni radicalmente incompatibili che implicano concezioni del mondo, della conoscenza e della mente fondamentalmente diverse. Il fatto che nessuna di esse riesca a eliminare le altre — che ognuna abbia un punto di forza che le altre non replicano — è il dato filosofico di partenza.

La risposta platonica ha il vantaggio dell'ovvietà: i teoremi matematici sembrano essere scoperti, non inventati. Euclide non ha *deciso* che i numeri primi sono infiniti; li ha dimostrati infiniti. Se fossero un'invenzione, si dovrebbe dire che prima di Euclide non vi erano infiniti numeri primi — ma questo sembra assurdo. Si aggiunge che la matematica sviluppata in isolamento dalla fisica si rivela poi adeguata a descrivere la struttura della realtà fisica: la geometria riemanniana, elaborata da Riemann nel 1854 come pura speculazione sulle varietà astratte, è risultata essere il linguaggio della relatività generale; la teoria degli operatori negli spazi di Hilbert, costruita senza alcun riferimento alla fisica, ha fornito il quadro formale della meccanica quantistica. Questa convergenza non si spiega facilmente se la matematica è una creazione culturale contingente; si spiega, almeno in apparenza, se i matematici esplorano un territorio di oggetti e strutture che esistono indipendentemente da loro.

Ma il prezzo filosofico del platonismo è esoso, e Paul Benacerraf lo ha messo a nudo con una chiarezza che ha orientato il dibattito per mezzo secolo. Nel saggio «Mathematical Truth» del 1973, Benacerraf formula il suo celebre dilemma: una teoria soddisfacente della verità matematica deve soddisfare simultaneamente due condizioni. La prima è che la semantica del discorso matematico sia uniforme con quella del discorso empirico — i termini singolari nelle asserzioni matematiche denotano oggetti esattamente come i termini singolari nelle asserzioni empiriche, e la struttura logica degli enunciati matematici è identica a quella degli enunciati non matematici. La seconda è che la conoscenza matematica sia inserita in una teoria epistemologica generale adeguata — si deve spiegare come esseri fisici e temporali, dotati di sistemi percettivi calibrati sulle interazioni causali con l'ambiente, possano venire a conoscere i fatti matematici.

La tensione tra queste due condizioni è, nella lettura di Benacerraf, quasi insormontabile per il platonismo. Se si adotta una concezione causale della conoscenza — secondo cui la conoscenza di p da parte di S richiede

un'appropriata relazione causale tra S e i fatti che rendono p vera — allora la conoscenza di oggetti astratti, causalmente inerti e collocati al di fuori dello spazio e del tempo, diventa del tutto misteriosa. Nessun platonista ha offerto una risposta davvero convincente a questa sfida. La risposta di Gödel, che postulava un'intuizione matematica di tipo quasi-percettivo capace di cogliere direttamente gli oggetti astratti, è insoddisfacente perché non spiega il meccanismo di tale intuizione né la sua affidabilità; le risposte di Maddy e di altri realisti naturalizzati, che cercano di fondare la conoscenza matematica sulla percezione degli insiemi concreti di oggetti fisici, si scontrano con l'irriducibilità dei concetti matematici più astratti alle situazioni percettive concrete.

L'alternativa nominalista — nella versione sviluppata da Hartry Field in *Science without Numbers* del 1980 — tenta di dissolvere il problema ontologico negando tout court l'esistenza degli oggetti astratti. La matematica, in questo quadro, è una grande finzione teoricamente conveniente: le asserzioni matematiche non descrivono un dominio reale ma sono strumenti per derivare conseguenze fisiche. La strategia di Field è tecnicamente ambiziosa. Egli dimostra che la meccanica newtoniana è nominalizzabile, ovvero riformulabile senza quantificazione su numeri o funzioni, impiegando esclusivamente relazioni tra eventi fisici concreti — relazioni spaziotemporali di congruenza e di betweenness, ispirate al programma di Hilbert per la geometria euclidea. La struttura formale della nominalizzazione è la seguente: Field introduce una teoria N che quantifica su regioni spaziotemporali concrete e postula relazioni primitive come la relazione quaternaria Cong(x, y, z, w) — «la distanza spaziotemporale tra x e y è uguale alla distanza tra z e w» — e la relazione Bet(x, y, z) — «y è tra x e z lungo una geodetica». Dimostra poi il teorema fondamentale: per ogni enunciato nominalista A appartenente al linguaggio di N (un enunciato che non quantifica su oggetti matematici), A è derivabile dalla congiunzione di N con la matematica standard se e soltanto se A è derivabile da N sola. In notazione: N + Matematica $\vdash$ A $\Leftrightarrow$ N $\vdash$ A, purché A sia nominalista. Questo è il risultato di conservatività: la matematica è conservativa rispetto alla fisica nominalista, non produce nuove conseguenze nominaliste, e può dunque essere trattata come uno strumento fittizio.

La critica più incisiva a questo programma è che la dimostrazione stessa del teorema di conservatività — nel meta-teorema che Field utilizza — ricorre a risorse matematiche non nominaliste. Se Field dimostra la conservatività usando la matematica, e poi invita il nominalista a trattare questa dimostrazione come finzione strumentale, ha semplicemente spostato il problema: il nominalista usa la matematica per stabilire un fatto sulla matematica, in una struttura che sembra circolare. Field risponde che il nominalista può accettare la dimostrazione come fiction senza impegno ontologico, ma questa risposta rende la posizione meno perspicua di quanto appaia, perché non chiarisce in che senso la dimostrazione sia epistemicamente affidabile se non la si considera vera.

Il costruttivismo brouweriano — che si esaminerà nella Parte Terza nel suo rapporto con le scienze cognitive — offre una terza via: gli oggetti matematici non esistono in un regno indipendente dalla mente né sono finzioni, ma sono costruzioni mentali effettuate nel tempo interiore della coscienza. Questo toglie il mistero dell'accessibilità epistemica (conosciamo gli oggetti matematici perché li

costruiamo), ma crea il problema della loro universalità: se i numeri naturali sono costruzioni della mia coscienza, perché i teoremi che li riguardano hanno la stessa struttura nelle coscienze di matematici vissuti in culture diverse e in epoche lontane?

Il panorama che emerge è quello di un problema genuinamente irrisolto, non per mancanza di ingegno filosofico ma per la struttura stessa della questione: essa richiede di connettere il livello della descrizione formale con il livello dell'esistenza e con il livello della cognizione, e questi tre livelli resistono ad ogni tentativo di riduzione a uno solo.

I.2 — Gottlob Frege e il logicismo: costruzione, crollo e permanenza

Il logicismo fregeano rappresenta il tentativo più ambizioso di risolvere il problema dell'ontologia matematica non per via ontologica ma per via logica. Se la matematica è riducibile alla logica pura — se i numeri sono oggetti logici definibili in termini puramente logici e i teoremi aritmetici sono teoremi logici — allora il loro status ontologico è quello degli oggetti logici e la loro conoscibilità è quella delle verità logiche: aprioristica, analitica, indipendente dall'esperienza. Il programma si svolge in tre tappe fondamentali che si distribuiscono tra il 1879 e il 1903: il *Begriffsschrift*, che introduce la notazione e il sistema deduttivo; i *Grundlagen der Arithmetik*, che fornisce l'analisi filosofica ; e i *Grundgesetze der Arithmetik*, che presenta il sistema formale completo.

La mossa decisiva dei *Grundlagen* consiste nel critica delle concezioni alternative del numero — il numero come proprietà delle cose, come insieme di unità, come libera creazione dell'intelletto — e nella proposta di una concezione propria: il numero è l'estensione di un concetto di secondo livello, ovvero la classe di tutti i concetti equinumerosi a un concetto dato. Due concetti F e G sono equinumerosi, secondo Frege, se e soltanto se esiste una relazione biunivoca che mette in corrispondenza gli F con i G — una definizione che evita di presupporre i numeri nella spiegazione dell'equinumerosità, usando al loro posto soltanto la nozione logica di biiezione. Formalmente, F è equinumeroso a G ($F \sim G$) se e soltanto se esiste una relazione R tale che:

$$\forall x(F(x) \rightarrow \exists !y(G(y) \wedge R(x,y))) \wedge \forall y(G(y) \rightarrow \exists !x(F(x) \wedge R(x,y)))$$

dove $\exists !$ denota il quantificatore «esiste esattamente uno». Il numero zero è poi la classe di tutti i concetti sotto cui non cade nessun oggetto; il numero uno è la classe di tutti i concetti sotto cui cade esattamente un oggetto; e in generale:

$$0 = \{F \mid \neg \exists x\, F(x)\}$$

$$1 = \{F \mid \exists x(F(x) \wedge \forall y(F(y) \rightarrow y = x))\}$$

dove $\{F \mid \varphi(F)\}$ denota l'estensione (la classe) del concetto φ. La definizione del successore è più delicata: n è il successore di m se e soltanto se esiste un concetto F e un oggetto x tale che F(x) e m è il numero del concetto «essere un F diverso da x» e n è il numero del concetto F. Formalmente:

$$n = succ(m) \leftrightarrow \exists F \exists x(F(x) \wedge m = \#\{y : F(y) \wedge y \neq x\} \wedge n = \#F)$$

dove #F denota il numero del concetto F — la classe di tutti i concetti equinumerosi a F.

L'intera costruzione presuppone il quinto assioma dei *Grundgesetze*, il controverso Legge Fondamentale V (Basic Law V, BLV), che afferma che due concetti hanno la stessa estensione se e soltanto se sono coestensivi:

BLV: $\varepsilon F = \varepsilon G \leftrightarrow \forall x(F(x) \leftrightarrow G(x))$

dove εF denota l'estensione del concetto F. Questo principio ha la struttura di un'astrazione: introduce nuovi oggetti — le estensioni — attraverso un criterio di identità che li individua attraverso una condizione sugli oggetti del livello inferiore. Frege era consapevole della problematicità del BLV: nei *Grundgesetze* lo caratterizza come «meno puramente logico degli altri assiomi» e avanza dubbi sulla sua sicurezza. La lettera di Bertrand Russell del giugno 1902, arrivata mentre il secondo volume era già in stampa, dimostrò che quei dubbi erano fondati nel modo più drammatico possibile.

Russell mostrò che il BLV implica una contraddizione. Si consideri la classe R di tutte le classi che non appartengono a se stesse:

$R = \varepsilon(\lambda x.\ x \notin x)$

Per il BLV, tale estensione esiste ed è un oggetto del dominio. Si pone allora la domanda: $R \in R$? L'applicazione diretta del BLV con $F = G = \lambda x.\ x \notin x$ e $x := R$ dà:

$R \in R \leftrightarrow R \notin R$

da cui segue immediatamente una contraddizione. L'annotazione di Frege in appendice al secondo volume — «Un geometra che trova il suo edificio crollare nel momento in cui ne ha completato le fondamenta si trova nella situazione più spiacevole possibile» — è uno dei momenti più tragicamente dignitosi nella storia della filosofia.

Frege tentò di emendare il BLV attraverso una versione indebolita (Basic Law V'), che restringe l'estensione dei concetti ai soli concetti che non si applicano a tutte le cose. Ma questa emenda — elaborata nell'appendice al secondo volume — fu mostrata inconsistente in presenza di certi altri assiomi del sistema da Lesniewski e poi da altri. Il programma logicista nella sua forma fregeana non sopravvisse alla crisi.

Ciò che sopravvive è di straordinaria fecondità. La notazione del *Begriffsschrift* — la prima presentazione rigorosa della logica del secondo ordine — ha trasformato la logica da disciplina syllogistica a sistema di deduzione formale. La distinzione tra senso e denotazione — di cui si tratterà nella Parte Seconda — è diventata il punto di partenza di tutta la filosofia del linguaggio contemporanea. E soprattutto, il logicismo neo-fregeano di Crispin Wright e Bob Hale — sviluppato a partire da *Frege's Conception of Numbers as Objects* (Wright, 1983) — ha mostrato che è possibile derivare l'aritmetica del secondo ordine non dal BLV ma dal solo Principio di Hume:

$$\mathrm{HP}: \#F = \#G \leftrightarrow F \sim G$$

senza ricorrere alla nozione problematica di estensione. Il Principio di Hume — che Frege aveva considerato un passo preliminare nella propria costruzione, preferendogli poi il BLV per ragioni di generalità — è un principio di astrazione logicamente consistente che introduce i numeri come oggetti il cui criterio di identità è la relazione di equinumerosità. Il risultato noto come Teorema di Frege afferma:

$$\mathrm{HP} \vdash_2 \mathrm{PA}_2$$

dove $\vdash_2$ indica la derivazione nel sistema della logica del secondo ordine e PA_2 denota l'aritmetica del secondo ordine (gli assiomi di Peano con la variabile insiemistica). In altri termini: dagli assiomi della logica del secondo ordine più il Principio di Hume, si possono derivare tutti i teoremi dell'aritmetica standard. Il logicismo in forma attenuata — più precisamente, in forma neo-fregeana — sopravvive dunque al crollo del BLV, sebbene la questione della natura epistemica e ontologica del Principio di Hume resti dibattuta.

Il dibattito contemporaneo verte in larga misura sullo status del Principio di Hume stesso: è un principio logico, come sosteneva Frege e come il neo-fregeismo implica? O è un principio sintetico a priori, nella tradizione kantiana? O è semplicemente una convenzione stipulativa che introduce i numeri per definizione, senza impegno ontologico reale? Wright e Hale sostengono che HP sia analitico nel senso che la sua verità può essere compresa attraverso la sola comprensione del significato dei suoi termini; Boolos controbatte che HP quantifica su oggetti — i numeri — e che questa quantificazione è ontologicamente impegnativa, non analitica. La questione è tuttora aperta.

I.3 — La teoria assiomatica degli insiemi: ZFC, la gerarchia cumulativa e i limiti del formalismo

L'alternativa che si è affermata come quadro fondazionale de facto della matematica del Novecento non è il logicismo né il nominalismo né il costruttivismo, ma la teoria degli insiemi assiomatica nel sistema Zermelo-Fraenkel con l'assioma della Scelta, universalmente noto con la sigla ZFC. Questo sistema non scioglie il problema ontologico — la sua semantica standard postula un universo V di tutti gli insiemi come dominio di quantificazione, il che presuppone una forma di platonismo — ma lo aggira dal punto di vista fondazionale, fornendo un sistema formale all'interno del quale la quasi totalità della matematica ordinaria può essere sviluppata in modo rigoroso senza richiedere una presa di posizione esplicita sulla natura degli oggetti matematici.

Gli assiomi di ZFC sono formulati nel linguaggio del primo ordine con un unico predicato primitivo di appartenenza $\in$. È utile esaminarne non soltanto la lista, ma anche la logica interna che li connette, perché questa logica rivela la strategia generale del sistema.

L'Assioma dell'Estensionalità stabilisce il criterio di identità per gli insiemi: due insiemi sono identici se e soltanto se hanno gli stessi elementi.

$$\forall x\ \forall y\ (\forall z\ (z \in x \leftrightarrow z \in y) \rightarrow x = y)$$

Questo assioma è l'analogo insiemistico del principio fregeano per i concetti: l'identità degli insiemi è determinata dal loro contenuto, non dalla loro modalità di presentazione. L'Assioma dell'Insieme Vuoto afferma l'esistenza di un insieme privo di elementi:

$$\exists x\ \forall y\ (y \notin x)$$

L'unicità è garantita dall'Estensionalità: l'insieme vuoto $\varnothing$ è unico. Gli Assiomi del Pairing e dell'Unione garantiscono la chiusura rispetto alle operazioni di costruzione:

Pairing: $\forall x\ \forall y\ \exists z\ \forall w\ (w \in z \leftrightarrow w = x \lor w = y)$ Unione: $\forall A\ \exists U\ \forall x\ (x \in U \leftrightarrow \exists y\ (y \in A \land x \in y))$

Lo Schema di Assiomi di Separazione è la risposta tecnica al paradosso di Russell: per ogni formula $\varphi(z)$ del linguaggio e ogni insieme A esiste un insieme B i cui elementi sono esattamente gli elementi di A che soddisfano φ:

$$\forall A\ \forall z_1 \ldots \forall z\ \ \exists B\ \forall x\ (x \in B \leftrightarrow x \in A \land \varphi(x, z_1, \ldots, z))$$

La restrizione cruciale è che si separa dall'insieme A preesistente, non dall'universo intero. Per definire l'analogo russelliano di R si dovrebbe separare dall'universo V intero, ma V non è un insieme nel sistema ZFC — è una classe propria. Il paradosso di Russell non si genera perché non si può applicare la Separazione all'universo intero.

L'Assioma dell'Insieme Potenza garantisce che per ogni insieme esista il suo insieme di parti:

$$\forall x\ \exists P\ \forall y\ (y \in P \leftrightarrow \forall z\ (z \in y \rightarrow z \in x))$$

Questo assioma, combinato con la Separazione, permette di costruire l'insieme $\mathbb{P}(A) = \{B : B \subseteq A\}$ per ogni insieme A. La sua importanza fondazionale è enorme: esso garantisce l'esistenza di insiemi di cardinalità molto più grande rispetto a quella degli insiemi da cui si parte, permettendo in particolare la costruzione del continuum $\mathbb{R}$ a partire da $\mathbb{N}$.

Lo Schema di Rimpiazzamento — introdotto da Fraenkel e Skolem per colmare una lacuna del sistema originale di Zermelo — afferma che l'immagine di un insieme attraverso una funzione definibile è essa stessa un insieme:

$$\forall A\ (\forall x \in A\ \exists! y\ \varphi(x,y)) \rightarrow \exists B\ \forall y\ (y \in B \leftrightarrow \exists x \in A\ \varphi(x,y))$$

Senza il Rimpiazzamento non sarebbe dimostrabile l'esistenza di insiemi come ω_1 (il primo ordinale non numerabile) o di insiemi come $\{\omega, P(\omega), P(P(\omega)), \ldots\}$ dove ω è l'insieme dei naturali, che non sono ottenibili dai soli assiomi di Zermelo.

L'Assioma dell'Infinito garantisce l'esistenza di un insieme induttivo:

$$\exists I\,(\varnothing \in I \wedge \forall x\,(x \in I \rightarrow x \cup \{x\} \in I))$$

Il più piccolo tale insieme è $\omega = \{\varnothing, \{\varnothing\}, \{\varnothing,\{\varnothing\}\}, ...\}$ — l'insieme dei numeri naturali nella costruzione di Von Neumann, in cui ogni naturale n è l'insieme di tutti i suoi predecessori: $n = \{0, 1, ..., n\text{-}1\}$.

L'Assioma di Fondazione esclude gli insiemi «mal fondati» — quelli che sarebbero membri di se stessi o che genererebbero catene discendenti infinite ($...\in z_2 \in z_1 \in z_0$):

$$\forall x\,(x \neq \varnothing \rightarrow \exists y\,(y \in x \wedge y \cap x = \varnothing))$$

Questo assioma garantisce che l'universo degli insiemi abbia la struttura gerarchica cumulativa, nota come gerarchia di Von Neumann: ogni insieme si trova in qualche strato V_α di una gerarchia indicizzata dagli ordinali. La costruzione della gerarchia procede come segue:

$V_0 = \varnothing$ $V_{\alpha+1} = \mathbb{P}(V_\alpha)$ per ogni ordinale α $V_\lambda = \cup_{\alpha<\lambda} V_\alpha$ per ogni ordinale limite λ

e infine $V = \cup\{a \in Ord\}\ V_a$. *Il rango di un insieme x è il minimo ordinale a tale che $x \subseteq V_a$, ovvero il minimo a tale che $x \in V_{\alpha+1}$.* L'universo degli insiemi V è interamente composto da insiemi puri — insiemi i cui elementi sono anch'essi insiemi, senza atomi non-insiemistici al fondo — e la struttura gerarchica garantisce che non vi siano cicli di appartenenza.

Infine, l'Assioma della Scelta afferma che per ogni famiglia di insiemi non vuoti esiste una funzione di scelta:

$$\forall A\,(\forall x \in A\ x \neq \varnothing \rightarrow \exists f\colon A \rightarrow \cup A\,(\forall x \in A\ f(x) \in x))$$

L'Assioma della Scelta è logicamente indipendente dagli altri assiomi di ZF — Gödel ne dimostrò la consistenza relativa nel 1938 costruendo il modello dell'universo costruibile L, e Cohen ne dimostrò l'indipendenza nel 1963 attraverso la tecnica del forcing — e ha conseguenze vaste e spesso controintuitive: il teorema di Banach-Tarski (una sfera può essere decomposta in un numero finito di parti e ricomposta in due sfere della stessa dimensione dell'originale), l'esistenza di un buon ordine su qualunque insieme, l'equivalenza tra l'esistenza di una base per ogni spazio vettoriale e l'AC.

I limiti di ZFC come fondamento della matematica si manifestano su due livelli distinti. Il primo è il livello tecnico delle proposizioni indipendenti: numerose questioni matematicamente significative non sono né dimostrabili né refutabili in ZFC, ammesso che ZFC sia consistente. L'esempio paradigmatico è l'Ipotesi del Continuum (CH) — la proposizione che asserisce che non esiste alcun cardinale strettamente compreso tra $\aleph_0$ (la cardinalità di ω) e $2^{\aleph_0}$ (la cardinalità del continuum). Gödel dimostrò nel 1938 la consistenza relativa di ZFC + CH mediante la costruzione del modello dell'universo costruibile:

$$L = \cup_{\alpha \in Ord} L_\alpha$$

dove $L_0 = \varnothing$, $L_{\alpha+1} = \mathrm{Def}(L_\alpha)$ (l'insieme delle parti di L_α definibili da formule del primo ordine con parametri in L_α), e $L_\lambda = \cup_{\alpha<\lambda} L_\alpha$ per i

limiti. Il risultato di Gödel è che CH è vera nel modello $\langle L, \in \rangle$, dove V = L (ogni insieme è costruibile). Cohen dimostrò nel 1963 — mediante la tecnica rivoluzionaria del forcing, che aggiunge nuovi insiemi «generici» a un modello base — la consistenza relativa di ZFC + ¬CH. Il forcing costruisce modelli in cui il continuum ha cardinalità $\aleph_2$, $\aleph_3$, $\aleph_\omega$, o quasi qualunque cardinale regolare: la cardinalità del continuum è intrinsecamente indeterminata nei termini di ZFC.

Questo risultato è epistemicamente destabilizzante se si ritiene che la matematica descrivi una realtà determinata. O il continuum ha o non ha cardinalità $\aleph_1$ — è difficile immaginare che entrambe le opzioni siano «vere» simultaneamente. Eppure ZFC non può decidere quale delle due sia il caso. La risposta platonica canonica di Gödel era che si dovranno trovare assiomi «evidenti» supplementari capaci di decidere CH; candidati come l'assioma di proiettiva determinatezza (PD), il Large Cardinal Axiom e varianti dell'assioma di forza (force axioms) sono stati proposti, ma nessuno ha acquisito lo status di evidenza che Gödel sognava. La risposta del pluralismo (Hamkins, 2012) è che non esiste un universo degli insiemi unico ma una pluralità di universi — il «multiverso degli insiemi» — tutti ugualmente legittimi come strutture formali; ma questa risposta sembra capitolare davanti all'indeterminazione piuttosto che risolverla.

Il secondo livello di limitazione è filosofico: la semantica standard di ZFC quantifica sull'universo V di tutti gli insiemi, ma questo universo non è esso stesso un insieme nel sistema ZFC — è una classe propria. Costruire un modello di ZFC richiede di operare in un meta-universo che è anch'esso un universo di insiemi, la cui consistenza non è dimostrabile in ZFC (per il Secondo Teorema di Incompletezza di Gödel). Il fondamento insiemistico di tutta la matematica richiede dunque un fondamento di se stesso che non fornisce. Questa circolarità non è banale: essa mostra che ZFC è uno strumento eccellente per produrre matematica ma non una risposta filosofica mente definitiva alla domanda «su che cosa si fonda la matematica?»

La teoria degli **assiomi dei grandi cardinali** — un programma di ricerca sviluppato principalmente da Jack Silver, William Mitchell, John Steel e Hugh Woodin — rappresenta il tentativo più ambizioso di rispondere alla indeterminatezza di ZFC attraverso l'aggiunta di assiomi «verticali» che estendono la gerarchia cumulativa verso ordinali di grandezza inimmaginabile. Un cardinale inaccessibile è un ordinale κ tale che è regolare (non cofinalità di ordinali più piccoli) e limite forte (per ogni $\alpha < \kappa$, $2^\alpha < \kappa$): la sua esistenza non è dimostrabile in ZFC ma è consistente con esso. I cardinali Mahlo, i cardinali misurabili, i cardinali di Woodin — ciascuno più forte del precedente nella gerarchia della forza consistenza — producono teorie via via più complete per la matematica dell'insieme. Il programma «Ultimate L» di Woodin mira a costruire un modello canonico delle assunzioni sui grandi cardinali che decida l'Ipotesi del Continuum — e i risultati preliminari suggeriscono che in tale modello CH sia vera. Ma la questione di se questi assiomi abbiano lo status di verità matematiche piuttosto che di ipotesi aggiuntive resta filosofica mente aperta e dipende in modo cruciale dalla concezione dell'ontologia matematica che si adotta.

I.4 — Lo strutturalismo matematico: Benacerraf, Shapiro e l'identità come posizione in una struttura

Il problema dell'indeterminazione dei referenti, formulato da Benacerraf nel saggio «What Numbers Could Not Be» del 1965, prende le mosse da un'osservazione a prima vista tecnica ma dalle conseguenze filosófiche profonde: esistono almeno due costruzioni insiemistiche dell'aritmetica ugualmente coerenti, ugualmente fedeli ai teoremi aritmetici, e tuttavia strutturalmente incompatibili nel senso che producono risposte opposte a domande insiemistiche determinate.

Nella costruzione di Von Neumann, ogni numero naturale n è identificato con l'insieme dei suoi predecessori:

$0 = \varnothing$, $1 = \{\varnothing\}$, $2 = \{\varnothing, \{\varnothing\}\}$, $3 = \{\varnothing, \{\varnothing\}, \{\varnothing,\{\varnothing\}\}\}$, ... $n = \{0, 1, ..., n\text{-}1\}$

In questa costruzione $3 \in 7$, perché $3 = \{0,1,2\}$ è un elemento di $7 = \{0,1,2,3,4,5,6\}$; la relazione di appartenenza insiemistica coincide con la relazione di minoranza tra numeri.

Nella costruzione di Zermelo, invece:

$0 = \varnothing$, $1 = \{\varnothing\}$, $2 = \{\{\varnothing\}\}$, $3 = \{\{\{\varnothing\}\}\}$, ... $n = \{n\text{-}1\}$

In questa costruzione $3 \notin 7$: il numero 3 è l'insieme $\{\{\{\varnothing\}\}\}$, mentre $7 = \{\{\{\{\{\{\{\varnothing\}\}\}\}\}\}\}$ non contiene $\{\{\{\varnothing\}\}\}$ come elemento diretto.

Entrambe le costruzioni soddisfano gli assiomi di Peano nella loro formulazione del secondo ordine:

PA_2: $\exists N\ \exists 0\ \exists S\ [0 \in N \land S: N \to N$ è iniettiva $\land\ 0 \notin S(N) \land\ \forall X(0 \in X \land\ \forall n(n \in X \to S(n) \in X) \to N \subseteq X)]$

dove N è il dominio dei naturali, 0 è lo zero e S è la funzione successore. Entrambe le costruzioni forniscono un insieme N, un elemento 0 e una funzione S iniettiva con $0 \notin S(N)$ tali che ogni sottoinsieme di N contenente 0 e chiuso per S sia uguale a N. Ma le domande «$2 \in 7$?» e «$\varnothing \in 3$?» ricevono risposte opposte nelle due costruzioni. Se i numeri naturali fossero identici agli insiemi di Von Neumann, la prima avrebbe risposta affermativa; se fossero gli insiemi di Zermelo, avrebbe risposta negativa.

Poiché non vi è alcuna ragione matematica per preferire l'una costruzione all'altra — entrambe sono categóricamente equivalenti come strutture aritmetiche — Benacerraf conclude che i numeri non possono essere identici né agli insiemi di Von Neumann né agli insiemi di Zermelo né ad alcun'altra costruzione insiemistica particolare. La conclusione è che i numeri non sono oggetti nel senso pieno del termine: non sono entità con un'identità intrinseca e proprietà indipendenti dalla struttura a cui appartengono. Essi sono, piuttosto, posizioni in strutture — posizioni nella progressione di Dedekind, una struttura $\langle N, 0, s\rangle$ che soddisfa gli assiomi di Peano. Ciò che conta è la struttura astratta; le particolari entità che la realizzano sono irrilevanti.

Questa intuizione è stata sistematizzata dallo strutturalismo matematico, nelle versioni complementari di Stewart Shapiro (*Philosophy of Mathematics: Structure and Ontology*, 1997) e di Michael Resnik (*Mathematics as a Science of Patterns*, 1997). Shapiro distingue tra sistemi — istanziazioni concrete o formali di strutture, come le costruzioni insiemistiche di Von Neumann e di Zermelo — e strutture stesse. La struttura aritmetica ω è un tipo di cui le diverse costruzioni insiemistiche, le sequenze di battiti cardiaci, le marcature su un nastro, i passi in un algoritmo, sono esemplari o istanze. Due sistemi esemplificano la stessa struttura se e soltanto se sono isomorfi: se esiste una biiezione $\psi: N_1 \to N_2$ tra i loro elementi che preserva tutte le relazioni rilevanti. Per la struttura aritmetica, due progressioni $\langle N_1, 0_1, s_1\rangle$ e $\langle N_2, 0_2, s_2\rangle$ sono isomorfe se e soltanto se esiste ψ tale che:

$$\psi(0_1) = 0_2 \;\forall n \in N_1: \psi(s_1(n)) = s_2(\psi(n))$$

Dedekind aveva dimostrato rigorosamente in *Was sind und was sollen die Zahlen?* del 1888 che ogni due progressioni semplicemente infinite sono isomorfe — che la struttura aritmetica è categorica (ogni due modelli sono isomorfi). Questo risultato è il fondamento técnico dello strutturalismo: se la struttura è unica a meno di isomorfismo, le sue posizioni — i «numeri» — non dipendono dal sistema particular che le realizza.

Shapiro distingue ulteriormente tra due varianti. Lo strutturalismo *ante rem* postula l'esistenza delle strutture come oggetti astratti genuini, indipendenti da qualunque istanziazione: la struttura ω esiste anche se non vi fosse alcun sistema concreto che la esemplificasse. Le strutture *ante rem* sono astratte — non spaziotemporalemente collocate né causalmente efficaci — ma sono oggetti con identità determinata dalla rete di relazioni che li costituiscono. La difficoltà epistemica è la stessa del platonismo tradizionale: come si conosce la struttura ω? Shapiro risponde attraverso la comprensione strutturale (*structural understanding*): il matematico conosce la struttura ω attraverso la sua caratterizzazione assiomatica a meno di isomorfismo. La struttura è, in qualche senso, la classe di equivalenza degli isomorfismi di sistemi, e la sua «identità» è costituita dall'invariante isomorfico.

Lo strutturalismo *in re*, sviluppato da Geoffrey Hellman in *Mathematics without Numbers* del 1989, rifiuta di postulare strutture astratte come entità indipendenti e interpreta le asserzioni matematiche come asserzioni modali condizionali. «Il numero 7 è primo» non asserisce che un certo oggetto abbia la proprietà di essere primo, ma che qualunque progressione semplicemente infinita possibile avrebbe nel settimo posto un elemento con tale proprietà:

$$\Box(\forall S)(PA_2(S) \to primo_7(S))$$

dove $PA_2(S)$ asserisce che S soddisfa gli assiomi di Peano del secondo ordine e $primo_7(S)$ asserisce che il settimo elemento di S è primo. La matematica non parla di oggetti ma di strutture possibili. Il prezzo è il ricorso alla modalità — al concetto di possibilità e necessità — come primitivo, e la questione se questa modalità strutturale sia ontologicamente meno impegnativa degli oggetti astratti è discussa.

La forza dello strutturalismo — nella sua versione *ante rem* come in quella modale — risiede nella sua capacità di spiegare l'indifferenza della matematica rispetto al substrato degli oggetti che la realizzano, il che è precisamente ciò che l'analisi di Benacerraf richiede. La sua debolezza risiede nella difficoltà di fornire un account del riferimento: quando un matematico parla del numero 7, a che cosa si riferisce esattamente? Lo strutturalismo risponde: alla posizione settima in qualunque progressione aritmetica. Ma questa risposta sembra presupporre che vi sia un'unica posizione settima — un'unica struttura ω — il che richiede, per lo strutturalismo *ante rem*, l'esistenza di tale struttura come entità astratta, ricadendo nella difficoltà epistemica originaria.

Il dibattito tra le diverse versioni dello strutturalismo e le posizioni alternative rivela la struttura profonda del problema ontologico matematico: non si tratta di scegliere tra oggetti astratti e nulla, ma di trovare il livello di descrizione appropriato — il livello al quale la matematica «parla di» qualcosa — senza cadere né nel platonismo epistemicamente misterioso né nel nominalismo che sacrifica troppo per evitare il mistero.

I.5 — Teoria delle categorie: morfismi, funtori e l'ontologia radicalmente relazionale

La teoria delle categorie, nata da una serie di articoli di Samuel Eilenberg e Saunders Mac Lane tra il 1942 e il 1945, costituisce non soltanto uno strumento tecnico di unificazione matematica — sebbene questo sia già di per sé considerevole — ma una riconfigurazione profonda del modo di pensare l'identità degli oggetti matematici. L'idea centrale è che, in matematica, gli oggetti non interessano in se stessi ma solo attraverso le trasformazioni che li connettono: un gruppo non è un insieme con una moltiplicazione; è un'entità che partecipa a morfismi di gruppo con altri gruppi, e la sua identità matemática è determinata dall'intera rete di questi morfismi. Questa intuizione — che lo strutturalismo filosofico di Shapiro e Resnik articola a livello filosofico — la teoria delle categorie la rende tecnica con una precisione che non lascia ambiguità.

Una **categoria** C è definita dal seguente dato:

una collezione ob(C) di oggetti (che può essere un insieme o una classe propria); per ogni coppia ordinata di oggetti (A, B), un insieme Hom_C(A, B) di morfismi da A a B — i morfismi f: A → B; per ogni terna di oggetti (A, B, C), un'operazione di composizione

$$\circ : \mathrm{Hom}(B, C) \times \mathrm{Hom}(A, B) \to \mathrm{Hom}(A, C)$$

che manda la coppia (g, f) con f: A → B e g: B → C nel morfismo composto g ∘ f: A → C; e per ogni oggetto A, un morfismo identità id_A ∈ Hom(A, A).

Questi dati devono soddisfare due assiomi. L'associatività della composizione:

$$h \circ (g \circ f) = (h \circ g) \circ f$$

per ogni f: A → B, g: B → C, h: C → D; e le leggi dell'identità:

$\mathrm{id}_B \circ f = f = f \circ \mathrm{id}_A$

per ogni $f: A \to B$.

La profondità filosofica di questa struttura emerge dalla caratteristica fondamentale che gli oggetti sono definiti esclusivamente attraverso i morfismi che li connettono: non hanno struttura interna descritta nel linguaggio della categoria. Un oggetto A di una categoria potrebbe essere un insieme, un gruppo, uno spazio topologico, una proposizione logica, un tipo computazionale — la struttura categoriale è indifferente a questa specificazione e opera al livello delle relazioni morfiche. Due oggetti A e B sono isomorfi ($A \cong B$) se e soltanto se esistono morfismi $f: A \to B$ e $g: B \to A$ tali che:

$g \circ f = \mathrm{id}_A$ e $f \circ g = \mathrm{id}_B$

Gli oggetti isomorfi sono, dal punto di vista della struttura categoriale, interscambiabili in ogni contesto: qualunque proprietà categoriale di A è una proprietà categoriale di B quando $A \cong B$. Questo è lo strutturalismo nel suo senso più rigoroso: le proprietà categorialmente significative sono invarianti per isomorfismo, e soltanto queste proprietà contano.

Un **funtore** covariante $F: C \to D$ tra categorie è un'operazione che preserva la struttura categoriale: assegna a ogni oggetto $A \in \mathrm{ob}(C)$ un oggetto $F(A) \in \mathrm{ob}(D)$; a ogni morfismo $f: A \to B$ in C un morfismo $F(f): F(A) \to F(B)$ in D; e soddisfa:

$F(g \circ f) = F(g) \circ F(f)$ per ogni $f: A \to B$, $g: B \to C$ $F(\mathrm{id}_A) = \mathrm{id}_{F(A)}$ per ogni $A \in \mathrm{ob}(C)$

Un funtore controvariante $G: C \to D$ inverte la direzione dei morfismi: se $f: A \to B$, allora $G(f): G(B) \to G(A)$. La controvarianza emerge naturalmente nelle situazioni in cui si ha a che fare con funzioni «inverse»: il funtore $\mathrm{Hom}(-, X)$ per un oggetto X fissato — che manda ogni oggetto A nell'insieme $\mathrm{Hom}(A, X)$ e ogni morfismo $f: A \to B$ nella funzione $\mathrm{Hom}(f, X): \mathrm{Hom}(B, X) \to \mathrm{Hom}(A, X)$ definita da $\mathrm{Hom}(f, X)(g) = g \circ f$ — è controvariante. La controvarianza riflette il fatto che le frecce vengono «girate»: una funzione da B a X si trasforma in una funzione da A a X pre-componendo con f.

Una **trasformazione naturale** $\eta: F \Rightarrow G$ tra due funtori paralleli $F, G: C \to D$ assegna a ogni oggetto $A \in \mathrm{ob}(C)$ un morfismo $\eta_A: F(A) \to G(A)$ in D (la componente di η in A) tale che per ogni morfismo $f: A \to B$ in C il diagramma

$F(A) \xrightarrow{\eta_A} G(A)$ | | $F(f)$ $G(f)$ | | $F(B) \xrightarrow{\eta_B} G(B)$

sia commutativo, ovvero: $G(f) \circ \eta_A = \eta_B \circ F(f)$. La nozione di trasformazione naturale fu introdotta da Eilenberg e Mac Lane precisamente per rendere precisa la distinzione intuitiva tra isomorfismi «naturali» — che non dipendono da scelte arbitrarie — e isomorfismi che dipendono da tali scelte.

Il caso classico dell'algebra lineare è illuminante. Per uno spazio vettoriale V su un campo k, esiste sempre un isomorfismo $V \to V^*$ (il duale), ma solo in dimensione finita — e questo isomorfismo non è naturale: dipende dalla scelta di

una base. Al contrario, l'isomorfismo $V \to V^{**}$ (il doppio duale) è naturale: è definito dalla funzione di valutazione $ev_V: V \to V^{**}$, $ev_V(v)(\varphi) = \varphi(v)$, e questa costruzione commuta con qualunque trasformazione lineare $T: V \to W$ nel senso che $ev_W \circ T = T^{**} \circ ev_V$, dove $T^{**}: V^{**} \to W^{**}$ è la trasposta iterata. La naturalità è esattamente la commutatività del diagramma della trasformazione naturale $ev: Id_{Vect_k} \Rightarrow (-)^{**}$.

Il **Lemma di Yoneda** è il risultato più profondo della teoria delle categorie elementare, e il suo significato filosofico merita un'analisi accurata. Sia C una categoria localmente piccola (ogni $Hom(A, B)$ è un insieme), $F: C^{op} \to Set$ un funtore controvariante dalla categoria opposta C^{op} alla categoria degli insiemi, e $A \in ob(C)$ un oggetto fissato. Il funtore rappresentabile in A è $Hom_C(-, A): C^{op} \to Set$, definito su oggetti da $Hom_C(-, A)(B) = Hom_C(B, A)$ e su morfismi $f: B \to C$ da $Hom_C(-, A)(f)(g) = g \circ f$ per $g: C \to A$. Il Lemma di Yoneda afferma che esiste una biiezione naturale:

$$Nat(Hom_C(-, A), F) \cong F(A)$$

In altri termini: l'insieme delle trasformazioni naturali dal funtore rappresentabile in A al funtore F è in biiezione naturale (naturale in A e in F) con l'insieme $F(A)$. La biiezione è data da:

$$\Phi: Nat(Hom_C(-, A), F) \to F(A), \quad \Phi(\eta) = \eta_A(id_A)$$

cioè: ogni trasformazione naturale $\eta: Hom(-, A) \Rightarrow F$ è univocamente determinata dall'elemento $\eta_A(id_A) \in F(A)$, l'immagine dell'identità $id_A \in Hom(A, A)$ sotto la componente $\eta_A: Hom(A, A) \to F(A)$. L'inversa manda $x \in F(A)$ nell'unica trasformazione naturale η^x tale che $\eta^x_B(f) = F(f)(x)$ per ogni $f: B \to A$.

Il significato filosofico del Lemma di Yoneda è questo: un oggetto A di una categoria è interamente determinato — a meno di isomorfismo — dalle sue relazioni con tutti gli altri oggetti, cioè dal funtore $Hom(-, A)$. L'ontologia categoriale è radicalmente relazionale: l'identità di un oggetto è la rete di relazioni morfiche che lo connette all'universo degli oggetti della categoria. Nessuna proprietà intrinseca — nessun «contenuto interno» — è rilevante al di là di questa rete. Il Lemma di Yoneda formalizza la tesi filosofica che lo strutturalismo aveva articolato in modo discorsivo: gli oggetti matematici non hanno essenza indipendentemente dalla struttura in cui sono implicati.

Si noti la portata di questo risultato. La piena embedding di Yoneda $\boldsymbol{y}: C \to Set^{C^{op}}$, $\boldsymbol{y}(A) = Hom_C(-, A)$, è un funtore pienamente fedele — cioè è iniettiva sugli oggetti a meno di isomorfismo e pienamente fedele sui morfismi: per ogni $A, B \in ob(C)$, la mappa $\boldsymbol{y}_{A,B}: Hom_C(A, B) \to Nat(Hom(-, A), Hom(-, B))$ è una biiezione. Questo significa che la categoria C si immerge fedelmente nella categoria dei presheaves $Set^{C^{op}}$: ogni informazione su oggetti e morfismi di C è catturata dai funtori rappresentabili. La categoria non perde nulla nel passaggio alla sua immagine di Yoneda — e questa immagine è una categoria di funzioni (presheaves) su C, cioè di oggetti descritti interamente dalle loro relazioni con tutti gli oggetti di C.

I.6 — Topos elementari e i fondamenti categoriali: logica interna e geometria del significato

La proposta fondazionale che dà la forma più compiuta all'alternativa categoriale alla teoria degli insiemi è la nozione di topos elementare, introdotta da F. William Lawvere e Myles Tierney a partire dal 1969-70. Un **topos elementare** è una categoria $\mathcal{E}$ dotata di struttura sufficiente per sviluppare all'interno di essa una logica e una matematica — non la logica classica della teoria degli insiemi, ma in generale la logica intuizionista della matematica costruttiva.

Le condizioni che definiscono un topos elementare sono quattro. Il topos deve avere un **oggetto terminale** 1: per ogni oggetto A esiste un unico morfismo $A \to 1$. Deve avere **prodotti finiti**: per ogni coppia A, B esiste il prodotto $A \times B$ con proiezioni $\pi_1: A \times B \to A$ e $\pi_2: A \times B \to B$ soddisfacenti la proprietà universale. Deve avere **oggetti esponenziali** B^A: per ogni coppia A, B esiste un oggetto esponenziale B^A con un morfismo di valutazione $ev: B^A \times A \to B$ tale che per ogni morfismo $f: C \times A \to B$ esiste un unico morfismo $\lambda f: C \to B^A$ con $ev \circ (\lambda f \times id_A) = f$. E deve avere un **classificatore di sottoggetti** Ω: esiste un oggetto Ω con un morfismo «vero» $true: 1 \to \Omega$ tale che per ogni monomorfismo $m: B \hookrightarrow A$ esiste un unico morfismo caratteristico $\chi_m: A \to \Omega$ che rende commutativo il quadrato di pullback:

$$\begin{array}{ccc} B & \longrightarrow & 1 \\ \downarrow & & \downarrow \text{true} \\ A & \xrightarrow{\chi_m} & \Omega \end{array}$$

ovvero: χ_m è la «funzione caratteristica» di B come sottooggetto di A. Il morfismo χ_m «classifica» il sottooggetto: esso manda gli elementi di A che appartengono a B al valore $true \in \Omega$ e gli altri a false (nel caso booleano). In un topos generale, l'oggetto Ω non è necessariamente booleano: può avere molti valori di «verità», corrispondenti ai diversi gradi di appartenenza al sottooggetto.

L'importanza filosofica di Ω risiede nel fatto che esso cattura la struttura logica del topos. Le operazioni logiche di congiunzione, disgiunzione, implicazione e negazione diventano morfismi dall'oggetto Ω:

$$\wedge: \Omega \times \Omega \to \Omega,\ \vee: \Omega \times \Omega \to \Omega,\ \Rightarrow: \Omega \times \Omega \to \Omega,\ \neg: \Omega \to \Omega$$

definiti attraverso le proprietà universali appropriate. Questi morfismi fanno di Ω una algebra di Heyting interna al topos — ovvero, le operazioni logiche soddisfano gli assiomi della logica intuizionista ma non necessariamente quelli della logica classica. In particolare, il principio del terzo escluso $p \vee \neg p = true$ non è necessariamente valido come morfismo $\Omega \to \Omega$.

Il topos Set degli insiemi ordinari — con la categoria degli insiemi e funzioni totali come topos elementare — ha logica classica: $\Omega = \{\bot, \top\} \cong 2$, e il terzo escluso vale. Ma il topos Sh(X) dei fasci su uno spazio topologico T non discreto ha logica intuizionista. Il classificatore di sottooggetti Ω nel topos Sh(X) è il fascio degli aperti: per ogni aperto $U \subseteq X$, $\Omega(U)$ è l'insieme degli aperti $V \subseteq U$. Ciò significa che i «valori di verità» in questo topos non sono semplicemente $\top$ e $\bot$, ma sono aperti di X — e la struttura logica riflette la topologia di X. Il principio del terzo escluso corrisponde, in questo contesto, alla condizione che

per ogni aperto U, $U \cup \mathrm{Int}(U^c) = X$ — una condizione che fallisce per qualunque spazio topologico non discreto con aperti che non siano né tutti né nessuno.

Questa connessione tra logica e geometria — tra struttura deduttiva e struttura topologica — è uno dei risultati più profondi e più belli della matematica del Novecento. Essa mostra che la logica non è monolitica: diversi topos realizzano diverse logiche, e la scelta della logica da adottare corrisponde alla scelta del topos (e dunque della geometria sottostante) in cui si lavora. Il platonismo degli insiemi (la logica classica) è soltanto un caso speciale — il caso del topos Set con la topologia discreta.

L'alternativa fondazionale dell'**Elementary Theory of the Category of Sets** (ETCS) di Lawvere (1964) ridescrive la categoria degli insiemi attraverso le sue proprietà categoriali, senza presupporre la nozione primitiva di appartenenza. Le funzioni (i morfismi) sono primitive, non derivate; gli «insiemi» (gli oggetti) sono caratterizzati dal loro comportamento rispetto alle funzioni. Gli «elementi» di un insieme A nella ETCS sono i morfismi $1 \to A$ dall'oggetto terminale: un elemento è un modo di selezionare un punto di A. La relazione $x \in A$ è rimpiazzata da $x: 1 \to A$ — un morfismo piuttosto che una relazione primitiva.

Come osserva Lawvere, questo spostamento rispecchia il modo in cui i matematici lavorano: in algebra, analisi e topologia, ciò che conta sono le trasformazioni tra strutture. L'abitudine di pensare agli insiemi come «contenitori di elementi» è un residuo delle origini della teoria degli insiemi nella tarda geometria del XIX secolo; l'abitudine categoriale di pensare agli insiemi come «nodi in una rete di funzioni» è più fedele alla pratica matematica effettiva. La filosofia che emerge è quella di un **strutturalismo funzionale**: le strutture matematiche sono essenzialmente reti di trasformazioni, non aggregati di oggetti.

I.7 — Teoria dei tipi di Martin-Löf, la corrispondenza di Curry-Howard e la triplice unità

La teoria dei tipi dipendente di Per Martin-Löf, elaborata in una serie di opere tra il 1975 e il 1984, rappresenta una delle sintesi intellettuali più profonde del Novecento: essa unifica la matematica costruttiva brouweriana, la logica deduttiva, la teoria della dimostrazione e la programmazione funzionale in un unico sistema formale — e lo fa attraverso una corrispondenza di straordinaria eleganza concettuale.

La **corrispondenza di Curry-Howard** — sviluppata nelle sue versioni originali da Haskell B. Curry negli anni Trenta e da William A. Howard nel 1969 (pubblicato nel 1980) — stabilisce un'identità strutturale tra sistemi logici e sistemi di tipi: le proposizioni logiche corrispondono ai tipi di dati; le dimostrazioni di proposizioni corrispondono ai programmi di quel tipo; la semplificazione di dimostrazioni (la normalizzazione delle prove) corrisponde all'esecuzione di programmi (la riduzione dei termini). Alcune corrispondenze fondamentali nella logica proposizionale:

$A \wedge B \leftrightarrow A \times B$ (tipo prodotto: una coppia di valori) $A \vee B \leftrightarrow A + B$ (tipo somma: un'alternativa tra due tipi) $A \to B \leftrightarrow A \to B$ (tipo funzione: dato un A, produce un B) $\bot \leftrightarrow \mathbb{0}$ (tipo vuoto: nessun programma lo abita) $\top \leftrightarrow \mathbb{1}$ (tipo unitario: un unico programma: $()$) $\neg A \leftrightarrow A \to \mathbb{0}$ (una funzione da A al tipo vuoto)

Nella teoria dei tipi dipendente di Martin-Löf, la corrispondenza si estende alla logica del primo ordine attraverso i **tipi dipendenti**, il cui nome descrive la loro caratteristica essenziale: il tipo di un'espressione può dipendere dal valore di un'altra espressione.

Il **tipo prodotto dipendente** $\Pi(x:A).\ B(x)$ è il tipo delle funzioni f tali che, per ogni termine a di tipo A, il termine f(a) ha tipo B(a), dove B è una famiglia di tipi parametrizzata da elementi di A. Quando B non dipende da x, $\Pi(x:A).\ B$ riduce al tipo funzione $A \to B$. Dal punto di vista logico, $\Pi(x:A).\ B(x)$ corrisponde alla proposizione $\forall x \in A.\ B(x)$: un termine f: $\Pi(x:A).\ B(x)$ è una dimostrazione costruttiva della proposizione universale — una procedura che, dato un abitante a: A (un oggetto del tipo A, cioè un testimone del fatto che A è abitata), produce una dimostrazione f(a): B(a) del fatto che B(a) vale.

Il **tipo somma dipendente** $\Sigma(x:A).\ B(x)$ è il tipo delle coppie dipendenti (a, b) dove a: A e b: B(a): il tipo del secondo componente dipende dal valore del primo componente. Dal punto di vista logico, $\Sigma(x:A).\ B(x)$ corrisponde alla proposizione $\exists x \in A.\ B(x)$: un termine (a, b): $\Sigma(x:A).\ B(x)$ è una dimostrazione costruttiva della proposizione esistenziale — con a: A il testimone e b: B(a) la prova che B(a) è soddisfatta per quel testimone.

Il **tipo identità** Id_A(a, b), per a, b: A, cattura la nozione di uguaglianza: un termine p: Id_A(a, b) è una «prova» che a e b sono uguali nel tipo A. L'unico costruttore è:

refl_a: Id_A(a, a)

la riflessività: ogni termine è uguale a se stesso. L'eliminatore del tipo identità — il **principio J** — ha la seguente forma:

$J: \Pi(A: \mathcal{U}).\ \Pi(C: \Pi(x\ y: A).\ Id_A(x,y) \to \mathcal{U}).\ (\Pi(a: A).\ C\ a\ a\ refl_a) \to \Pi(x\ y: A).\ \Pi(p: Id_A(x,y)).\ C\ x\ y\ p$

Questo principio afferma che per dimostrare una proprietà C delle prove di uguaglianza p: Id_A(x, y), è sufficiente dimostrarla nel caso x = y e p = refl_x. Il principio J cattura l'idea che l'uguaglianza ha un comportamento uniforme: tutto ciò che vale per la riflessività vale per qualunque prova di uguaglianza.

Nella teoria standard di Martin-Löf, il principio J implica che due prove di uguaglianza tra i medesimi termini sono esse stesse uguali — il principio di Unicità delle Prove di Identità (UIP):

$UIP: \Pi(a\ b: A).\ \Pi(p\ q: Id_A(a,b)).\ Id_\{Id_A(a,b)\}(p,q)$

Con UIP, la struttura del tipo identità è «discreta»: non vi sono prove di uguaglianza non banali, e la struttura è quella di un insiemismo ordinario.

I.8 — Homotopy Type Theory: l'assioma di univalenza e il fondamento omotopico

Il progetto della Homotopy Type Theory (HoTT) nasce dall'osservazione di Vladimir Voevodsky, intorno al 2006-2009, che la struttura del tipo identità nella teoria di Martin-Löf senza l'assioma UIP si comporta in modo sorprendentemente analogo agli spazi topologici e ai loro gruppoidi fondamentali. L'interpretazione omotopica dei tipi, sviluppata indipendentemente da Awodey e Warren e da Voevodsky, ha aperto una nuova via di fondazione della matematica esposta nel «HoTT Book» del 2013.

L'interpretazione omotopica associa a ogni tipo A una struttura omotopica:

i termini a: A sono i **punti** dello spazio topologico A; i termini p: Id_A(a, b) sono i **cammini** (paths) da a a b; i termini H: Id_{Id_A(a,b)}(p, q) sono le **omotopie** tra cammini; e così via per ogni livello superiore: le omotopie tra omotopie, ecc.

In questa prospettiva, l'assenza di UIP non è un difetto ma una ricchezza: apre la possibilità che il tipo Id_A(a, b) abbia più di un abitante — prove di uguaglianza non riducibili alla riflessività, corrispondenti, nell'interpretazione topologica, a cammini non-banali.

La concatenazione di cammini p: Id_A(a, b) e q: Id_A(b, c) produce $p \cdot q$: Id_A(a, c) (transitività dell'uguaglianza); l'inversione di p produce p^{-1}: Id_A(b, a) (simmetria); la composizione è associativa a meno di omotopia superiore. Questa struttura forma, per ogni tipo A, uno ∞-groupoide — una struttura categoriale debole con morfismi a tutti i livelli, tutti invertibili a meno di morfismi del livello successivo. La classificazione dei tipi per complessità omotopica procede attraverso la nozione di n-troncatura.

Un tipo A è una **(-2)-troncatura** (tipo contraibile) se ha esattamente un abitante a meno di uguaglianza: $\exists$(a:A). Π(x:A). Id_A(a, x). Un tipo è una **(-1)-troncatura** (h-proposizione) se ha al più un abitante: Π(x y: A). Id_A(x, y). Un tipo è una **0-troncatura** (insieme omotopico) se tutte le sue prove di uguaglianza sono uguali — Π(x y: A). Π(p q: Id_A(x,y)). Id_{Id_A(x,y)}(p,q) — cioè se soddisfa UIP. Questo è il livello degli insiemi ordinari. Un tipo è una **1-troncatura** (gruppoide omotopico) se può avere prove di uguaglianza non banali ma tutte le omotopie tra cammini sono banali. E così via.

L'**assioma di univalenza** di Voevodsky è formulato come:

UA: $(A \simeq B) \simeq (A =_{\mathcal{U}} B)$

dove $A \simeq B$ denota il tipo delle equivalenze omotopiche tra A e B — funzioni f: $A \to B$ che ammettono una quasi-inversa g: $B \to A$ con omotopie $g \circ f \sim id_A$ e $f \circ g \sim id_B$ — e $A =_{\mathcal{U}} B$ è il tipo identità nell'universo $\mathcal{U}$. Informalmente: due tipi sono «uguali» se e soltanto se sono equivalenti omotopicamente. L'assioma di univalenza afferma che l'uguaglianza di tipi nell'universo coincide con la loro equivalenza strutturale.

La conseguenza filosofica più immediata dell'UA è il **principio di invarianza per equivalenza**: qualunque proprietà P: $\mathcal{U} \to \mathcal{U}$ definibile nel linguaggio della teoria dei tipi deve essere invariante per equivalenza omotopica. Se $A \simeq B$ e P(A) vale, allora P(B) vale. Questo è il principio strutturalista formulato in modo computazionalmente verificabile: le proprietà matematiche non dipendono dall'identità numerica degli oggetti ma soltanto dalla loro struttura omotopica.

L'assioma UA ha conseguenze matematiche concrete di grande portata. Una delle più immediate è che i tipi Fin(2) = {0, 1} e $\mathbb{B}$ = {true, false} — che sono equivalenti tramite la biiezione che manda 0 a false e 1 a true — diventano uguali nell'universo per UA. Qualunque proprietà valida di Fin(2) vale dunque automaticamente di $\mathbb{B}$ senza bisogno di ulteriori dimostrazioni. Nell'approccio insiemistico ZFC, Fin(2) e $\mathbb{B}$ sono insiemi con elementi diversi; il trasferimento di proprietà richiede una dimostrazione esplicita che rispetti la struttura. In HoTT, l'equivalenza produce un percorso nell'universo, e il trasporto di proprietà lungo questo percorso è il meccanismo base della teoria.

I **tipi induttivi superiori** (HIT) estendono le possibilità della teoria: si possono definire tipi non soltanto specificando costruttori di punti, ma anche costruttori di cammini e omotopie superiori. Il cerchio S^1 — che non può essere definito come tipo induttivo ordinario, perché ha gruppo fondamentale $\pi_1(S^1) = \mathbb{Z}$ non banale — è definito come HIT con:

base: S^1 (un costruttore di punto) loop: Id_{S^1}(base, base) (un costruttore di cammino non banale)

Una conseguenza immediata è che si può dimostrare, all'interno di HoTT, il teorema di Hopf fondamentale: $\pi_1(S^1) \simeq \mathbb{Z}$. Ogni intero $n \in \mathbb{Z}$ corrisponde a una classe di omotopia di cammini base → base in S^1 — il cammino che percorre il cerchio n volte (con segno che indica il verso). Questo risultato, che nella topologia algebrica classica richiede macchine omologiche sofisticate, emerge in HoTT come conseguenza diretta della struttura del tipo identità su S^1 e del principio J.

La connessione tra algebra (i gruppi di omotopia) e logica (i tipi identità) è il cuore concettuale di HoTT. Essa mostra che la distinzione tra «matematica» e «logica» — cruciale per il logicismo fregeano — diventa permeabile a un livello sufficientemente astratto: la logica è essa stessa matematica omotopica, e la matematica omotopica è logica dei tipi. Il fondamento omotopico della matematica non è né platonico né nominalista né costruttivista in senso stretto: è strutturalista nel senso più radicale, in quanto identifica l'uguaglianza con l'equivalenza strutturale e tratta tutta la matematica come lo studio di strutture invarianti per equivalenza.

I.9 — L'efficacia irragionevole della matematica: strutture formali e realtà fisica

La questione più perturbante che l'ontologia matematica deve affrontare non è soltanto il problema interno di Benacerraf, ma la domanda esterna formulata da Eugene Wigner nel celebre articolo «The Unreasonable Effectiveness of

Mathematics in the Natural Sciences» del 1960: perché le strutture matematiche, sviluppate spesso senza alcuna considerazione per le applicazioni fisiche, risultano poi straordinariamente adatte a descrivere il comportamento della natura?

I casi emblematici sono numerosi e ben noti, ma il loro cumulare non smette di stupire. La geometria riemanniana, elaborata da Riemann nel discorso inaugurale del 1854 come pura speculazione su varietà n-dimensionali con tensore metrico variabile, fornisce il linguaggio della relatività generale. Una varietà riemanniana $\langle M, g\rangle$ è uno spazio differenziabile M dotato di un tensore metrico g — un campo di forme bilineari simmetriche definite positive sui tangenti — che definisce lunghezze e angoli localmente. Il tensore di Riemann, definito dalla connessione di Levi-Civita ∇ compatibile con g e priva di torsione:

$$R(X,Y)Z = \nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z - \nabla_{[X,Y]} Z$$

misura la curvatura intrinseca della varietà — la misura in cui il trasporto parallelo di un vettore attorno a un loop lo ruota. Il tensore di Ricci $\mathrm{Ric}(X,Y) = \mathrm{tr}(Z \mapsto R(Z,X)Y)$ e la curvatura scalare $R = \mathrm{tr}\,\mathrm{Ric}$ sono le contrazioni del tensore di Riemann che compaiono nelle equazioni di campo di Einstein:

$$G_{\mu\nu} + \Lambda g_{\mu\nu} = (8\pi G/c^4)\, T_{\mu\nu}$$

dove $G_{\mu\nu} = \mathrm{Ric}_{\mu\nu} - \tfrac{1}{2} R g_{\mu\nu}$ è il tensore di Einstein, Λ è la costante cosmologica, G è la costante di Newton, c è la velocità della luce, e $T_{\mu\nu}$ è il tensore energia-impulso. La struttura matematica elaborata per ragioni puramente geometriche sessant'anni prima era, in un senso che Einstein non poteva anticipare, il linguaggio della gravitazione.

Analogamente, le strutture di gruppo — studiate da Galois nell'ambito della teoria delle equazioni algebriche, sviluppate da Lie per le trasformazioni continue, astratte da Dedekind e altri nel quadro dell'algebra moderna — costituiscono l'impalcatura formale di tutta la fisica delle particelle elementari attraverso la teoria dei campi di gauge. Il Modello Standard della fisica delle particelle è basato sul gruppo di simmetria $\mathrm{SU}(3) \times \mathrm{SU}(2) \times \mathrm{U}(1)$, e i bosoni di gauge (i mediatori delle forze) sono connessioni su fibrati principali con struttura di gruppo uguale a questo gruppo di gauge. La struttura algebrica che determina quali particelle esistono e come interagiscono è precisamente la struttura dei gruppi di Lie semisemplici e delle loro rappresentazioni.

La teoria degli operatori negli spazi di Hilbert — sviluppata da Hilbert, von Neumann e altri a partire dagli anni Venti come estensione astratta dell'analisi funzionale, senza alcun riferimento alla fisica — fornisce il quadro formale della meccanica quantistica. Uno spazio di Hilbert $\mathcal{H}$ è uno spazio vettoriale complesso con prodotto interno $\langle \cdot \mid \cdot \rangle$ che è completo rispetto alla norma $\|\psi\| = \sqrt{\langle \psi \mid \psi \rangle}$. Le osservabili fisiche corrispondono a operatori autoaggiunti $A = A^\dagger$ su $\mathcal{H}$; gli stati fisici corrispondono a vettori unitari $\psi \in \mathcal{H}$ (o più precisamente a raggi — classi di equivalenza di vettori che differiscono per un fattore di fase). Il valore atteso di un'osservabile A nello stato ψ è:

$$\langle A \rangle_\psi = \langle \psi \mid A \mid \psi \rangle$$

e la misurazione di A produce con probabilità $|\langle a|\psi\rangle|^2$ l'autovalore a corrispondente all'autovettore $|a\rangle$ — la regola di Born. L'equazione di Schrödinger:

$$i\hbar\,\partial|\psi\rangle/\partial t = H|\psi\rangle$$

dove H è l'operatore hamiltoniano (l'energia del sistema), è un'equazione differenziale nello spazio di Hilbert $\mathcal{H}$ — una struttura che Hilbert aveva sviluppato per motivi puramente matematici e che si è rivelata essere il linguaggio della meccanica quantistica.

L'interpretazione di questo fenomeno ha diviso i filosofi della matematica in posizioni nette. Il **realismo strutturale** — nelle versioni di John Worrall, Steven French e James Ladyman — sostiene che l'efficacia della matematica rivela che la struttura del mondo fisico è essa stessa matematica: ciò che la fisica scopre non è la natura intrinseca degli oggetti fisici ma la struttura delle relazioni tra essi, che coincide con struttura matematica. Le particelle elementari non sono «fatte di» qualcosa di non-matematico; esse sono nodi in una rete di relazioni di simmetria che è precisamente la struttura dei gruppi di gauge. L'«accordo irragionevole» tra matematica e fisica non è irragionevole: matematica e fisica hanno lo stesso oggetto — le strutture — e convergono perché descrivono la medesima cosa.

La risposta nominalista è che la convergenza è il risultato di una selezione: tra le infinite strutture matematicamente possibili, i fisici hanno selezionato quelle che si adattano ai dati osservativi. La maggior parte delle strutture matematiche non ha applicazioni fisiche; quelle che le hanno sono quelle che i fisici hanno trovato utili dopo averle cercate. Ma questa risposta non scioglie del tutto il mistero: il fatto che strutture sviluppate in isolamento assoluto dalla fisica — la geometria riemanniana, gli spazi di Hilbert — si rivelino poi adeguate a descrivere fenomeni fisici scoperti indipendentemente, suggerisce qualcosa di più di una selezione retrospettiva. Il matematico che studia le varietà riemanniane non sa che Einstein le userà; il fisico che cerca un linguaggio per la gravitazione non conosce il lavoro di Riemann. La convergenza avviene tra stranieri.

La risposta del costruttivismo cognitivo di Lakoff e Núñez — che la matematica è radicata nelle strutture dell'esperienza corporea e l'efficacia riflette la storia evolutiva dei suoi creatori — ha difficoltà a spiegare le strutture matematiche che sembrano procedere a distanza considerevole da qualunque esperienza corporea possibile. La teoria dei motivi di Grothendieck, i ∞-gruppoidi di HoTT, la teoria delle forme automorfiche — sono strutture la cui rilevanza fisica non è immediata ma che si rivela in contesti sempre nuovi. Se esse fossero soltanto proiezioni dell'esperienza corporea, sarebbe difficile capire perché si rivelino efficaci per descrivere il mondo sub-atomico, che è radicalmente inaccessibile all'esperienza corporea.

La domanda rimane aperta — e questa apertura non è una sconfitta filosofica ma il sintomo di una profondità genuina. Essa costituisce la cornice entro cui le riflessioni sulle quattro parti di questa monografia trovano il loro orizzonte più vasto: il problema del significato che abita le strutture formali non è separabile dal problema del rapporto tra le strutture formali e la realtà fisica che esse descrivono. Entrambi i problemi si chiudono su questa domanda: che cosa

permette a una struttura — simbolica, formale, matematica — di raggiungere ciò che sta al di là di se stessa? È in questo spazio di domande irrisolte che si aprono le questioni della filosofia del linguaggio, delle scienze cognitive e dell'intelligenza artificiale che si esaminano nelle parti successive.

I.10 — Il predicativismo e la stratificazione fondazionale: Weyl, Feferman e il limite di Schütte-Feferman

Il predicativismo rappresenta una posizione fondazionale che occupa uno spazio concettuale distinto tanto dal platonismo quanto dal nominalismo, e la cui chiarificazione illumina aspetti cruciali del rapporto tra struttura formale e contenuto matematico che i fondamentalismi estremi tendono a oscurare. La sua genealogia intellettuale risale alle preoccupazioni di Henri Poincaré riguardo alle definizioni impredicative — definizioni che presuppongono ciò che definiscono, come la definizione del minimo di un insieme attraverso una quantificazione su tutti i reali che include il minimo stesso. Poincaré denominò «circoli viziosi» (*cercles vicieux*) queste strutture definitorie e suggerì che il paradosso di Russell fosse precisamente un caso di tale circolarità: la classe R di tutte le classi che non appartengono a se stesse è definita attraverso una quantificazione su tutte le classi, inclusa R stessa.

Bertrand Russell rispose a questa diagnosi costruendo la teoria dei tipi ramificata, in cui ogni proposta si colloca in un livello di tipo determinato e le quantificazioni sono ristrette a oggetti di tipo inferiore. La ramificazione impedisce le auto-applicazioni impredicative: non è possibile definire un oggetto di tipo n attraverso una quantificazione su oggetti di tipo n o superiore, perché quella quantificazione presupporrebbe già come dato ciò che si sta definendo. Il costo della ramificazione è enorme: la matematica classica risulta largamente ineseguibile senza l'assioma di riduttibilità — assioma che afferma l'esistenza, per ogni funzione proposizionale ramificata di ogni livello, di una funzione equivalente al livello più basso. Russell e Whitehead introdussero tale assioma nei *Principia Mathematica*, ma con disagio evidente: come notò Frank Ramsey, l'assioma di riduttibilità è una petizione di principio che reintroduce esattamente l'impredicatività che la ramificazione intendeva eliminare, e la sua verità è contingente piuttosto che logica.

Hermann Weyl ha sviluppato, nel trattato *Das Kontinuum* del 1918, il programma predicativista più rigoroso e matematicamente elaborato della prima fase. Weyl accetta la totalità dei numeri naturali come dato primitivo — come la struttura aritmetica originaria a cui la mente ha accesso diretto attraverso l'intuizione della successione — ma nega che si possa quantificare liberamente sulle totalità di oggetti di ordine superiore (insiemi di naturali, insiemi di insiemi di naturali) trattandole come domini completi disponibili nella loro interezza. La distinzione cruciale è quella tra operazioni costruttive, che generano nuovi oggetti a partire da oggetti già dati, e quantificazioni impredicative, che presuppongono domini già completati come sfondo di quantificazione. Il sistema di Weyl permette di definire un insieme di naturali A soltanto attraverso predicati che quantificano esclusivamente sui naturali, non su insiemi di naturali: ogni definizione deve

essere riferita al livello inferiore già disponibile, senza «guardare in avanti» verso il dominio degli insiemi di cui A farebbe parte.

Weyl dimostra che un sottoinsieme sorprendentemente ricco dell'analisi reale classica è sviluppabile all'interno di questi vincoli: i teoremi fondamentali del calcolo differenziale e integrale, la teoria delle serie convergenti, una porzione sostanziale della teoria della misura di Lebesgue. Tuttavia, alcune strutture centrali dell'analisi classica rimangono fuori portata — l'estremo superiore di un insieme limitato di reali (che richiede quantificazione sull'insieme di tutti i reali, ovvero sul continuum come totalità), la costruzione del completamento di uno spazio metrico arbitrario, certi teoremi di punto fisso come il teorema di Brouwer. Queste dimostrazioni richiedono di trattare il continuum non come un processo di costruzione in corso ma come una totalità già data e percorribile nella sua interezza — e questa è precisamente l'assunzione che il predicativismo rifiuta.

Il programma predicativista ha ricevuto nel secondo Novecento una precisazione matematica di grande eleganza attraverso i lavori di Solomon Feferman e Kurt Schütte. La domanda fondamentale che Feferman pone è: quale è il limite esatto della matematica sviluppabile predicativamente? La risposta è caratterizzata dal sistema della «predicatività data dai naturali», che può essere descritta formalmente attraverso una gerarchia transfinita di sistemi la cui solidità cresce lungo la gerarchia degli ordinali costruttivi. Ogni sistema nella gerarchia può quantificare su gli insiemi definibili nel sistema precedente, e la totalità dei sistemi predicativamente ammissibili è indicizzata da ordinali che sono essi stessi costruibili predicativamente. Il limite di questa gerarchia ascendente — l'ordinale Γ_0, denominato il limite di Feferman-Schütte — è il più piccolo ordinale che non è accessibile attraverso le operazioni di costruzione degli ordinali che il predicativismo ammette.

Formalmente, Γ_0 può essere caratterizzato come il più piccolo punto fisso della funzione di Veblen $\varphi_\alpha(0)$: $\Gamma_0 = \min\{\alpha : \varphi_\alpha(0) = \alpha\}$, dove φ_α è la funzione di Veblen a due argomenti che diagonalizza sulla gerarchia degli enumeratori degli ordinali critici. Questa definizione, pur nella sua complessità tecnica, cattura precisamente il punto in cui la costruzione predicativa si ferma: al di là di Γ_0, la definizione degli ordinali richiede operazioni che presuppongono come dato qualcosa che non può essere costruito predicativamente. La forza dimostrativa del sistema predicativista — misurata in termini degli ordinali il cui buon ordine è dimostrabile nel sistema — è esattamente Γ_0.

La teoria delle «matematiche inverse» (*Reverse Mathematics*) sviluppata da Harvey Friedman e Steve Simpson ha fornito uno strumento per misurare con precisione il contenuto predicativo dei teoremi matematici. Il sistema ACA_0 (*Arithmetic Comprehension Axiom*) — un sistema del secondo ordine che permette la comprensione di insiemi di naturali definibili da formule aritmetiche, ovvero da formule che quantificano soltanto sui naturali — cattura l'essenza del predicativismo di Weyl nel linguaggio formale dell'aritmetica del secondo ordine. La forza di ACA_0 è proprio Γ_0, e una vasta classe di teoremi matematici classici è stata dimostrata essere equivalente ad ACA_0 sopra una base debole RCA_0 — il che significa che questi teoremi sono dimostrabili predicativamente e non richiedono strumenti impredicativi.

Il significato filosofico del predicativismo per il problema dell'ontologia matematica è duplice e di notevole profondità. In primo luogo, esso mostra che la matematica non è un monolite indiviso: esiste una stratificazione interna che distingue livelli di impegno ontologico richiesti da diversi risultati. Un teorema dimostrabile predicativamente ha uno status ontologico diverso da uno che richiede impredicatività: il primo non presuppone totalità infinite come domini di quantificazione completi, il secondo sì. Il predicativismo introduce quindi una «topografia» del fondazionale che il platonismo e il nominalismo tendono entrambi ad appiattire. In secondo luogo, il predicativismo rimodella la domanda dell'ontologia matematica in modo produttivo: invece di chiedere se gli oggetti matematici esistano in astratto, chiede quali operazioni costruttive sono legittime e quali domande matematiche sono determinatamente rispondibili con i mezzi costruttivi disponibili. La domanda non è «il continuum esiste?» ma «quali proprietà del continuum sono determinatamente decidibili con i mezzi che abbiamo costruttivamente a disposizione?» — una domanda che apre un programma di ricerca invece di presupporre la risposta.

Il dibattito tra predicativismo e impredicativismo non riguarda soltanto i fondamenti tecnici della matematica: esso è la versione precisa del problema, che attraversa tutta questa monografia, di come la struttura formale possa raggiungere il contenuto che la trascende. La matematica predicativa si ferma dove non può costruire il proprio dominio di quantificazione; la matematica impredicativa salto al di là, presupponendo domini come già dati. Questo salto è esattamente il punto in cui la struttura tocca il proprio limite — dove il sistema formale richiederebbe, per dimostrare la propria completezza, risorse che non possiede internamente.

PARTE SECONDA: FILOSOFIA DEL LINGUAGGIO E SEMANTICA FORMALE

II.1 — Senso, denotazione e la stratificazione irriducibile del significato

Il salto dalla Parte Prima alla Parte Seconda non è un cambiamento di argomento ma un cambiamento di angolatura: si passa dalla domanda su che cosa sono le strutture matematiche alla domanda su come le espressioni linguistiche — e le strutture formali che le rappresentano — raggiungono i loro oggetti. La matematica, come si è visto, si confronta con il problema di come la struttura formale si connetta a una realtà che la trascende; la filosofia del linguaggio pone la medesima domanda in modo più diretto, perché il linguaggio ordinario ha referenti che non sono oggetti astratti in un regno platonico ma individui, eventi, proprietà di un mondo concreto. Che cosa permette a un'espressione di raggiungere il proprio referente? E quando due espressioni raggiungono il medesimo referente per vie diverse, sono semanticamente equivalenti o differiscono per qualcosa di essenziale?

Queste domande ricevono la loro prima formulazione rigorosa nell'articolo che Gottlob Frege pubblica nel 1892 con il titolo «Über Sinn und Bedeutung» — un

testo di quaranta pagine che ha trasformato la semantica più di qualunque altro contributo prima o dopo di esso. L'ingresso nel problema avviene attraverso il fenomeno delle identità informative: enunciati della forma «a = b» che non sono tautologie ridondanti ma ampliativi, che comunicano qualcosa di genuino sul mondo e hanno un valore cognitivo che «a = a» non possiede.

Il caso fregeano è astronomico nella sua semplicità pedagogica: «l'Astro della Sera» e «l'Astro del Mattino» sono due espressioni che denotano entrambe il pianeta Venere — lo stesso punto nello spazio, la stessa massa planetaria, lo stesso corpo celeste. Se il significato di un'espressione singolare fosse identico al suo referente, i due sintagmi sarebbero semanticamente equivalenti — intercambiabili in ogni contesto senza alterazione del valore semantico. Ma non è così. Si confrontino:

(1) L'Astro della Sera è l'Astro del Mattino. (2) L'Astro della Sera è l'Astro della Sera.

L'enunciato (1) è stato stabilito empiricamente — dagli osservatori babilonesi, o dai Pitagorici secondo alcune fonti — ed è una scoperta astronomica genuina. L'enunciato (2) è una tautologia, conoscibile a priori senza osservazione alcuna. Se il significato delle espressioni fosse il loro comune referente (Venere), i due enunciati avrebbero lo stesso contenuto semantico — sarebbero veri esattamente nelle stesse circostanze, avrebbero lo stesso valore cognitivo. Ma non lo hanno. La differenza di valore cognitivo richiede, per Frege, la postulazione di un livello di contenuto distinto dalla denotazione: il **senso** (Sinn), il modo in cui l'espressione presenta o determina il referente.

Il senso di «l'Astro della Sera» è qualcosa come «il corpo celeste visibile in occidente dopo il tramonto»; il senso di «l'Astro del Mattino» è «il corpo celeste visibile in oriente prima dell'alba». I due sensi sono differenti — presentano il medesimo oggetto attraverso modalità di determinazione distinte — e questa differenza di senso spiega la differenza di valore cognitivo tra (1) e (2). L'enunciato (1) è informativo perché asserisce l'identità di oggetti presentati attraverso sensi diversi; stabilire questa identità richiede verifica empirica perché le due modalità di presentazione non si identificano analiticamente.

La struttura semantica fregeana si estende dal livello dei termini singolari al livello degli enunciati interi con coerenza sistematica. La **denotazione** (Bedeutung) di un enunciato completo è il suo valore di verità — il Vero o il Falso, concepiti come oggetti logici dello stesso tipo logico degli oggetti individuali. Il **senso** di un enunciato è il **pensiero** (Gedanke) che esso esprime: un contenuto proposizionale oggettivo, non dipendente dalla psicologia di nessun parlante particular, afferrabile da più soggetti e suscettibile di essere vero o falso. L'oggettività dei pensieri è fondamentale per l'epistemologia fregeana: i teoremi matematici sono veri perché i pensieri che esprimono corrispondono a fatti logici oggettivi, non perché qualcuno li pensi in un certo modo.

Il problema dei **contesti intensionali** — i contesti «opachi» o «obliqui» in cui la sostituzione di espressioni co-denotative non preserva il valore di verità dell'enunciato — è la prova più diretta della necessità di un livello semantico oltre la denotazione. Si consideri:

(3) Giovanna crede che l'Astro della Sera sia luminoso. (4) Giovanna crede che l'Astro del Mattino sia luminoso.

Se Giovanna ignora che l'Astro della Sera e l'Astro del Mattino sono lo stesso corpo celeste — il che è storicamente possibile — (3) può essere vera e (4) falsa. Ma se il valore semantico delle espressioni nei contesti di credenza fosse la loro denotazione ordinaria (Venere), le due espressioni dovrebbero essere interscambiabili senza effetto sul valore di verità dell'enunciato. La non-interscambiabilità rivela che in contesti obliqui le espressioni non si riferiscono ai propri referenti ordinari ma ai propri sensi ordinari: in (3) e (4), «l'Astro della Sera» e «l'Astro del Mattino» denotano i rispettivi sensi, che sono diversi anche se le denotazioni ordinarie coincidono.

Questa soluzione genera una gerarchia semantica potenzialmente infinita. Se «crede che φ» produce un contesto obliquo in cui le espressioni di φ si riferiscono ai propri sensi ordinari, cosa accade nei doppi contesti obliqui — «crede che [crede che φ]»? Le espressioni di φ in questo contesto si devono riferire ai propri sensi indiretti (i sensi dei propri sensi ordinari), che sono a loro volta distinti dai sensi ordinari. Ogni livello di annidamento introduce un nuovo livello semantico. Michael Dummett, nella ricostruzione monumentale della filosofia di Frege (*Frege: Philosophy of Language*, 1973), ha tentato di sistematizzare questa gerarchia, ma senza raggiungere un accordo unanime sulla sua adeguatezza — e senza dissipare il sospetto che la teoria fregeana dei contesti obliqui sia vittima di una regressione mal governata.

Il punto filosofico di più ampia portata che emerge dalla distinzione senso-denotazione è la struttura a due livelli del significato. Il livello della denotazione è il livello dell'estensione — dell'oggetto nel mondo che l'espressione raggiunge. Il livello del senso è il livello dell'intensione — del modo in cui l'oggetto è presentato, del contenuto cognitivo dell'espressione. Questa distinzione non è soltanto una risposta tecnica al problema delle identità informative; è la struttura portante di tutta la filosofia del linguaggio successiva. Le teorie semantiche che si costruiranno nei decenni seguenti — la semantica dei mondi possibili, la grammatica di Montague, la semantica distribuzionale — sono tutte, in modi diversi, tentativi di fare justice a questa duplice dimensione del significato: la dimensione in cui un'espressione raggiunge il mondo e la dimensione in cui lo raggiunge in un certo modo.

II.2 — La teoria delle descrizioni di Russell: la forma logica nascosta e l'analisi del non-essere

La risposta di Bertrand Russell al problema delle identità informative e dei contesti opachi diverge radicalmente da quella fregeana: invece di introdurre un secondo livello semantico accanto alla denotazione, Russell nega che le descrizioni definite siano genuini termini singolari dotati di referente, e le analizza come costruzioni quantificazionali complesse — quantificatori camuffati dalla superficie grammaticale. La teoria è esposta nel celebre articolo «On Denoting» del 1905, che Willard Van Orman Quine ha definito «un paradigma della filosofia».

Il problema che motiva la teoria di Russell è più ampio di quello fregeano: esso riguarda non soltanto le identità informative ma l'intero dominio delle descrizioni che sembrano riferirsi a oggetti che non esistono. L'enunciato «il Re attuale di Francia è calvo» ha una superficie grammaticale di predicazione — un soggetto singolare e un predicato — ma il soggetto non ha referente (la Francia è una repubblica). Se la funzione semantica di un termine singolare è denotare un oggetto, e il termine singolare non denota alcun oggetto, allora o l'enunciato è privo di valore di verità (Frege: la denotazione è assente, dunque il valore di verità manca) o è falso per ragione di presupposizione fallita. Ma entrambe le opzioni sono problematiche: la prima rende una gran parte del discorso ordinario semanticamente difettosa; la seconda non riesce a spiegare perché «il Re di Francia non è calvo» sembri avere lo stesso problema di «il Re di Francia è calvo».

L'analisi russelliana dell'enunciato «il Re attuale di Francia è calvo» decompone la predicazione apparente nella congiunzione di tre asserzioni distinte: che esiste almeno un individuo che è attualmente Re di Francia; che esiste al più un tale individuo (unicità); che ogni tale individuo è calvo. Formalmente:

$$\exists x[RF(x) \wedge \forall y(RF(y) \rightarrow y = x) \wedge Calvo(x)]$$

dove RF(x) abbrevia «x è attualmente Re di Francia». Poiché la prima clausola è falsa — non vi è alcun Re di Francia — l'intera congiunzione è falsa. Non priva di valore di verità, non né vera né falsa: semplicemente falsa, perché la condizione esistenziale non è soddisfatta. Questo è il guadagno principale rispetto a Frege: si elimina la necessità di postulare denotazioni vuote o oggetti inesistenti.

La struttura dell'analisi rivela l'idea fondamentale: la forma grammaticale superficiale — «il Re di Francia è calvo» come predicazione su un soggetto — è fuorviante rispetto alla **forma logica** dell'enunciato, che è invece una complessa asserzione quantificazionale. La distinzione tra forma grammaticale e forma logica — tra la struttura della lingua come appare in superficie e la struttura che deve emergere per una semantica rigorosa — diventa uno dei temi centrali della filosofia del linguaggio novecentesca. Wittgenstein nel *Tractatus* ne fa una premessa metodologica fondamentale: la proposizione nasconde la propria forma logica sotto la veste della grammatica ordinaria, e il compito della filosofia è rivelare la forma logica nascosta. Chomsky nel programma della grammatica generativa propone una versione linguistica della distinzione — la struttura superficiale contro la struttura profonda — che ne è chiaramente debitrice, sebbene i fini e i metodi differiscano.

La descrizione definita «ι x φ(x)» — l'unico x tale che φ(x) — non è un termine singolare primitivo ma un operatore definito contestualmente attraverso l'analisi quantificazionale:

$$\psi(\iota x\, \varphi(x)) =_{def} \exists x(\varphi(x) \wedge \forall y(\varphi(y) \rightarrow y = x) \wedge \psi(x))$$

Questo è il principio di eliminazione contestuale delle descrizioni: non si definisce il referente della descrizione, ma si elimina la descrizione dall'enunciato in cui compare, sostituendola con la sua analisi quantificazionale. Le descrizioni non hanno significato isolato; hanno soltanto uso in contesto.

Una conseguenza importante dell'analisi russelliana è il trattamento delle descrizioni nei contesti modali e nelle negazioni. La frase «il Re di Francia non è calvo» è ambigua tra due letture, a seconda della portata della negazione rispetto all'operatore descrittivo:

Portata ampia: $\neg \exists x[RF(x) \wedge \forall y(RF(y) \rightarrow y = x) \wedge Calvo(x)]$ Portata ristretta: $\exists x[RF(x) \wedge \forall y(RF(y) \rightarrow y = x) \wedge \neg Calvo(x)]$

Nella prima lettura, l'enunciato asserisce che non è il caso che esista un unico Re di Francia calvo — il che è vero, perché non esiste alcun Re di Francia. Nella seconda lettura, asserisce che esiste un unico Re di Francia e che non è calvo — il che è falso per le stesse ragioni della versione affermativa. Russell spiega le intuizioni conflittuali di Frege sulla negazione (che a volte sembra terziare tra vera, falsa e indeterminata) attraverso questa distinzione di portata: non vi è indeterminatezza semantica, ma soltanto ambiguità sintattica nella portata degli operatori.

L'articolo di Keith Donnellan «Reference and Definite Descriptions» del 1966 ha introdotto una complicazione significativa nella dibattito. Donnellan osserva che le descrizioni definite possono essere usate in due modi distinti. L'**uso attributivo** usa la descrizione per descrivere un tipo di individuo: «chiunque abbia ucciso Smith è pazzo» — si asserisce qualcosa di chiunque soddisfi la descrizione, senza avere in mente un individuo specifico. L'analisi russelliana coglie questo uso: l'enunciato è vero se e soltanto se l'unico individuo che ha ucciso Smith è pazzo. L'**uso referenziale** usa invece la descrizione per riferirsi a un individuo specifico che il parlante ha in mente, indipendentemente dal fatto che la descrizione si applichi letteralmente a quell'individuo: «il galantuomo con il martini in mano», detto di qualcuno che in realtà ha succo di pompelmo nel bicchiere, è una descrizione usata referenzialmente — il parlante si riferisce a un individuo specifico e usa la descrizione soltanto come dispositivo per attirare l'attenzione su di lui. In quest'uso, la teoria di Russell sembra fallire: anche se la descrizione non si applica letteralmente, il riferimento riesce.

Il dibattito sulla portata della distinzione di Donnellan — se si tratti di un fenomeno semantico che richiede una revisione dell'analisi russelliana, o di un fenomeno puramente pragmatico spiegabile nei termini dell'intenzione comunicativa del parlante senza modificare la semantica — non si è concluso. Kripke, in «Speaker's Reference and Semantic Reference» (1977), sostiene che la distinzione è pragmatica: il significato semantico delle descrizioni è quello quantificazionale russelliano, ma i parlanti possono usarle con riferimento a un individuo specifico tramite l'intenzione comunicativa, e questa intenzione appartiene alla pragmatica del discorso, non alla semantica delle espressioni. Donnellan risponde che la distinzione ha ripercussioni sulle condizioni di verità e non può essere relegata alla sola pragmatica. La questione resta un punto di tensione tra la semantica del significato delle espressioni e la pragmatica del loro uso in contesto — una tensione che attraversa tutta la filosofia del linguaggio successiva.

II.3 — Semantica dei mondi possibili: Kripke, la rigidità e la revisione della mappa modale

La nozione di mondo possibile — utilizzata informalmente dai logici modali medievali e riapparsa nella filosofia di Leibniz come strumento per analizzare la modalità — riceve la sua formalizzazione matematica completa in una serie di articoli di Saul Kripke tra il 1959 e il 1963, e la sua elaborazione filosofica più profonda nelle conferenze poi pubblicate come *Naming and Necessity* (1972, 1980). La semantica dei mondi possibili trasforma la logica modale da un sistema puramente sintattico di assiomi e regole in un sistema con una semantica composizionale ben definita, e apre al contempo una revisione radicale della mappa delle distinzioni epistemiche e metafisiche che la tradizione analitica dava per scontate.

Un **frame modale di Kripke** è una coppia $\langle W, R \rangle$ dove W è un insieme non vuoto di mondi possibili e $R \subseteq W \times W$ è una relazione binaria di accessibilità tra mondi — intuitivamente, wRw' significa «da w è possibile che w' sia il caso». Un **modello** $M = \langle W, R, V \rangle$ aggiunge una funzione di valutazione V che assegna a ogni formula atomica p e ogni mondo w un valore $V(p, w) \in \{0, 1\}$. Le condizioni di verità per i connettivi booleani sono le condizioni standard, mondo per mondo. Per gli operatori modali:

$M, w \vDash \Box\varphi \leftrightarrow \forall w' \in W\ (wRw' \rightarrow M, w' \vDash \varphi)$ $M, w \vDash \Diamond\varphi \leftrightarrow \exists w' \in W\ (wRw' \wedge M, w' \vDash \varphi)$

Così $\Box\varphi$ — «necessariamente φ» — è vera in w se e soltanto se φ è vera in tutti i mondi accessibili da w; $\Diamond\varphi$ — «possibilmente φ» — è vera in w se e soltanto se φ è vera in almeno un mondo accessibile.

Le proprietà della relazione R corrispondono ad assiomi modali con interpretazioni filosófiche precise. Se R è riflessiva ($\forall w: wRw$), si ottiene l'assioma T: $\Box\varphi \rightarrow \varphi$ — ciò che è necessariamente vero è vero nel mondo attuale. Se R è transitiva ($wRw' \wedge w'Rw'' \rightarrow wRw''$), si ottiene l'assioma 4: $\Box\varphi \rightarrow \Box\Box\varphi$ — ciò che è necessariamente vero è necessariamente necessariamente vero. Se R è euclidea ($wRw' \wedge wRw'' \rightarrow w'Rw''$), si ottiene l'assioma 5: $\Diamond\varphi \rightarrow \Box\Diamond\varphi$. La combinazione di T, 4 e 5 equivale a richiedere che R sia una relazione di equivalenza (riflessiva, transitiva, simmetrica), e dà il sistema modale S5 — la logica modale «massimale» tra quelle normali standard — in cui ogni mondo è accessibile da ogni altro e la necessità è assoluta: $\Box\varphi \leftrightarrow \neg\Diamond\neg\varphi$, e $\Box\varphi$ e $\Diamond\varphi$ hanno il medesimo valore di verità in tutti i mondi simultaneamente.

L'importanza filosofica più duratura di *Naming and Necessity* non risiede però nella semantica formale della logica modale, che avrebbe potuto rimanere un esercizio tecnico, bensì nella revisione radicale della mappa concettuale che le tradizioni analitica e kantiana avevano ereditato. Tre distinzioni erano state sistematicamente confuse — trattate come se fossero tre facce dello stesso fenomeno:

la distinzione **metafisica** tra necessario e contingente: un enunciato è necessario se vero in tutti i mondi possibili, contingente se vero in alcuni e falso in altri;

la distinzione **semantica** tra analitico e sintetico: un enunciato è analitico se vero in virtù del significato dei suoi termini (la sua negazione è contraddittoria), sintetico se la sua verità dipende da fatti che vanno oltre il significato;

la distinzione **epistemica** tra a priori e a posteriori: un enunciato è conoscibile a priori se la sua verità può essere accertata senza ricorrere all'esperienza empirica, a posteriori se richiede verifica empirica.

Kripke dissolve la presunta identità tra queste tre distinzioni attraverso la nozione di **designatore rigido**: un termine che denota il medesimo individuo in tutti i mondi possibili in cui quell'individuo esiste. I nomi propri («Aristotele», «Venere») e i termini di tipo naturale («acqua», «tigre», «oro») sono designatori rigidi; le descrizioni definite («il discepolo di Platone che scrisse la Metafisica») in generale non lo sono, perché denotano in mondi diversi individui diversi — l'individuo che soddisfa la descrizione in quel mondo.

La rigidità dei nomi propri implica che le identità tra nomi propri — quando vere — sono necessariamente vere. Se «Espero» e «Fosforo» designano entrambi rigidamente il pianeta Venere nel mondo attuale, allora in ogni mondo possibile in cui Espero esiste e Fosforo esiste, essi designano lo stesso pianeta. L'enunciato «Espero è Fosforo», se vero, è vero in tutti i mondi possibili — è necessariamente vero. Ma esso è stato scoperto a posteriori, attraverso osservazione astronomica. Si ottiene così la categoria delle **verità necessarie a posteriori**, che la tradizione aveva ritenuto vuota.

L'argomento merita di essere esaminato nella sua struttura dettagliata. La necessità dell'identità non è una conseguenza di alcuna definizione; essa segue dalla rigidità dei designatori. Se d_1 e d_2 sono designatori rigidi dello stesso oggetto x nel mondo attuale, allora per ogni mondo possibile w: d_1 denota x in w (se x esiste in w) e d_2 denota x in w (se x esiste in w). Non esiste alcun mondo possibile in cui d_1 denoti un oggetto y e d_2 denoti un oggetto z con $y \neq z$, perché sia d_1 che d_2 denotano rigidamente x. Dunque in ogni mondo in cui entrambi i designatori hanno referente, essi hanno lo stesso referente: l'identità di Espero e Fosforo non potrebbe essere altrimenti, perché non è metafisicamente possibile che i due designatori rigidi co-denotanti nel mondo attuale denotino oggetti distinti in qualche altro mondo.

La necessità è metafisica, non epistemica. Epistémicamente, prima dell'osservazione astronomica che ha identificato Espero con Fosforo, era aperta per noi la possibilità che si trattasse di corpi celesti diversi — questa era una possibilità epistemica, una situazione che per quanto ne sapevamo avremmo potuto trovarci. Ma una possibilità epistemica non è una possibilità metafisica: il fatto che qualcosa fosse aperto per la nostra conoscenza non implica che fosse metafisicamente possibile. L'aprioricità e l'aposterioricità riguardano l'accessibilità epistemica; la necessità e la contingenza riguardano la struttura metafisica dei mondi possibili. Le due dimensioni sono indipendenti.

Kripke identifica anche la categoria simmetrica delle **verità contingenti a priori**, attraverso l'uso delle descrizioni come «fissatori di referente». Si immagini che si stipuli che «Napole» denoti rigidamente il fondatore attuale dell'azienda X — chiunque egli sia. In questo uso, «Napole è il fondatore di X» è conoscibile a

priori (in virtù della stipulazione) ma è contingente (il fondatore attuale avrebbe potuto non fondare l'azienda). La necessità a posteriori e la contingenza a priori mostrano insieme che le tre distinzioni filosófiche — metafisica, semantica, epistemica — formano una mappa a nove caselle, di cui Kant e la tradizione avevano esplorato soltanto tre.

Il programma kripkeano viene completato dalla tesi sull'identità della mente e del cervello. Se gli stati mentali e gli stati neurali sono identici — se il dolore è identico all'attivazione delle fibre C, come il fisicalismo identitario sostiene — allora, per la necessità dell'identità tra designatori rigidi, questa identità sarebbe necessaria: non ci sarebbero mondi possibili in cui il dolore esiste senza l'attivazione delle fibre C, né mondi in cui le fibre C si attivano senza dolore. Ma questo sembra falso: sembra perfettamente possibile immaginare un organismo con le fibre C attivate ma senza dolore, o un organismo con dolore realizzato da una struttura fisica diversa. La risposta del fisicalista — che questa è soltanto una possibilità epistemica — è disponibile, ma richiede uno sforzo argomentativo considerevole per essere resa plausibile. L'argomento di Kripke contro il fisicalismo identitario di tipo è uno degli argomenti più discussi nella filosofia della mente contemporanea.

II.4 — Realismo modale, mondi possibili e la semantica proposizionale intensionale

David Lewis ha elaborato, in *On the Plurality of Worlds* del 1986, una risposta al problema ontologico dei mondi possibili che costituisce la posizione filosofica più radicalmente coerente e più ontologicamente provocatoria nel panorama del realismo modale. La posizione dominante prima di Lewis — il realismo modale moderato di Robert Adams, Alvin Plantinga e altri — tratta i mondi possibili come entità astratte concrete: massime proposizioni consistenti, o stati di cose massimali possibili, che non sono universi fisici ma oggetti astratti simili a proposizioni o insiemi.

Lewis rifiuta questa concezione come oscura e filosofica mente insoddisfacente: che cosa renderebbe concrete le «proposizioni massimali» se non i fatti che descrivono? Egli propone invece il **realismo modale concreto**: i mondi possibili sono universi spaziotemporali concreti, ugualmente reali al mondo in cui abitiamo, privi di connessioni causali tra loro. L'aggettivo «attuale» (actual) è un indicale che varia riferimento da mondo a mondo — esattamente come «qui» varia da luogo a luogo e «ora» varia da tempo a tempo. Quando diciamo «il mondo attuale esiste», affermiamo semplicemente, dall'interno del nostro mondo, che il nostro mondo esiste; non stiamo attribuendo una proprietà ontologica intrinseca che lo distingua dagli altri mondi possibili. Ogni abitante di ogni mondo chiama «attuale» il proprio mondo.

La riduzione della modalità alla quantificazione su entità concrete è il guadagno teorico fondamentale del realismo di Lewis: «è necessario che p» significa «in tutti i mondi, p» — per ogni universo concreto w, p è vera in w. «È possibile che p» significa «in qualche mondo, p» — esiste un universo concreto w in cui p è vera. La modalità non è una nozione primitiva e incompresa; è quantificazione su un

dominio di entità della stessa categoria ontologica delle entità concrete del mondo attuale. I controfattuali — «se A fosse accaduto, B sarebbe accaduto» — ricevono la loro analisi in termini di similarità tra mondi: «se A fosse il caso, allora B sarebbe il caso» è vera nel mondo attuale @ se e soltanto se i mondi più simili a @ in cui A è vera sono mondi in cui anche B è vera. La relazione di similarità è rappresentata da una struttura di sfere centrate su @: le sfere più piccole contengono i mondi più simili a @, e le sfere più grandi i mondi meno simili. L'enunciato controfattuale è vero se e soltanto se ogni mondo nella sfere più piccola in cui A è vera è anche un mondo in cui B è vera.

Il significato linguistico, nel quadro di Lewis, viene analizzato in termini di funzioni caratteristiche su mondi possibili. La **proposizione** espressa da un enunciato è l'insieme dei mondi possibili in cui l'enunciato è vero — o equivalentemente, la funzione caratteristica da W a {0, 1} che vale 1 esattamente nei mondi in cui l'enunciato è vero. Questa concezione ha eleganza formale: la sinonimia è l'identità di proposizioni; la necessità è la proposizione universale $\{w \in W : 1\} = W$; l'impossibilità è la proposizione vuota $\varnothing$; l'implicazione è l'inclusione di proposizioni.

I limiti della concezione proposizionale intensionale emergono però quando si presta attenzione alla **iperintensionalità** — ai contesti in cui la sostituzione di proposizioni logicamente equivalenti non preserva il valore di verità. Se la proposizione di «7 è primo» e la proposizione di «la somma degli angoli interni di un triangolo euclideo è 180°» sono entrambe necessariamente vere — vere in tutti i mondi possibili — allora nella semantica lewisiana esse sono la medesima proposizione, l'insieme W di tutti i mondi. Ma «Giulio crede che 7 sia primo» e «Giulio crede che la somma degli angoli interni di un triangolo euclideo sia 180°» non hanno lo stesso valore di verità in tutti i contesti: Giulio potrebbe credere il primo senza credere il secondo (se non conosce la geometria euclidea). La semantica proposizionale intensionale — anche quella lewisiana — non riesce a catturare le distinzioni semantiche nei contesti iperintensionali.

Le risposte a questa difficoltà hanno preso due direzioni principali. La prima è l'introduzione di strutture proposizionali più fini — proposizioni strutturate, che non sono insiemi di mondi ma oggetti strutturati che riflettono la struttura sintattica degli enunciati che le esprimono. La proposizione espressa da «7 è primo» è l'ordinata ⟨primo, 7⟩ — una coppia composta dalla proprietà di essere primo e dall'individuo 7 — anziché un insieme di mondi. La seconda direzione è l'introduzione di **mondi impossibili** — mondi in cui valgono contraddizioni o falsità matematiche — per distinguere proposizioni logicamente equivalenti ma «impossibilmente» diverse. In un mondo impossibile in cui 7 non è primo, la proposizione «7 è primo» è falsa, anche se in tutti i mondi ordinari (non impossibili) è vera. Questa seconda direzione è filosofica mente controversa perché richiede di accordare realtà (o quasi-realtà) a mondi che violano le leggi della logica — ma è tecnicamente utile per trattare le credenze inconsistenti e i contesti iperintensionali.

II.4bis — Semantica bidimensionale: carattere, contenuto e la geometria della necessità a posteriori

La distinzione kripkeana tra necessità metafisica e aprioricità epistemica — tra le verità vere in tutti i mondi possibili e le verità conoscibili senza esperienza — ha prodotto, nelle generazioni successive, il tentativo di costruire un quadro semantico capace di ospitare entrambe le dimensioni all'interno di un unico sistema formale. Questo tentativo prende la forma della **semantica bidimensionale**, sviluppata in versioni complementari da David Kaplan nel contesto delle espressioni indicali (*Demonstratives*, 1989), da Martin Davies e Lloyd Humberstone (1980) nel contesto della logica modale, e da David Chalmers nell'ambito del dibattito sulla coscienza e la necessità a posteriori. La semantica bidimensionale non è una proposta unitaria ma una famiglia di approcci che condividono la struttura fondamentale: il significato di un'espressione è una funzione a due argomenti invece di uno, e le due dimensioni corrispondono rispettivamente alla dimensione epistemica e alla dimensione metafisica del significato.

Il punto di partenza è la semantica degli **indicali** di Kaplan. Un'espressione come «io», «qui», «ora», «questo» non denota lo stesso individuo in ogni contesto di uso: «io» denota il parlante, «qui» denota il luogo, «ora» denota il momento dell'enunciazione. La semantica degli indicali richiede di distinguere due livelli di contenuto. Il **carattere** (*character*) di un'espressione è una funzione dai contesti di enunciazione alle estensioni: il carattere di «io» è la funzione che manda ogni contesto $c = \langle$agente, luogo, tempo, mondo$\rangle$ all'agente di c. Il **contenuto** (*content*) di un'espressione in un contesto è il valore del carattere per quel contesto — l'individuo, la proposizione o la proprietà che l'espressione esprime in quel particolare contesto. Per «io» nel contesto in cui sono il parlante, il contenuto è me stesso. La composizionalità semantica opera sul contenuto, non sul carattere: la proposizione espressa da una frase è una funzione dalle circostanze di valutazione (mondi possibili) ai valori di verità, e questa funzione si costruisce compositivamente dai contenuti delle parti.

La struttura bidimensionale emerge quando si riconosce che sia il carattere sia il contenuto possono essere valutati in diversi «mondi». Il carattere di un'espressione è stabile: è la stessa funzione indipendentemente da come il mondo avrebbe potuto essere. Ma il mondo in cui viene valutata la proposizione espressa può variare: «io sono qui» è vera nel mondo attuale, ma potrebbe essere falsa in mondi in cui il parlante si trovasse altrove. Kaplan distingue allora tra la modalità che opera sul contesto — la variazione di chi sta parlando, dove e quando — e la modalità che opera sulle circostanze di valutazione. Questa distinzione è cruciale: una proposizione espressa da un indicale può essere necessariamente vera *rispetto alle circostanze di valutazione* (vera in ogni mondo possibile) senza che il carattere che la produce sia necessariamente applicabile (un agente diverso produrrebbe un'enunciazione diversa). «Io esisto», se enunciata da me, esprime una proposizione che è necessariamente vera (ogni mondo in cui mi trovo è un mondo in cui esisto) — ma il carattere dell'enunciato non è necessario, perché un agente diverso produrrebbe proposizioni diverse.

Davies e Humberstone hanno formalizzato la struttura bidimensionale attraverso due operatori modali distinti: l'operatore □ di necessità ordinaria — che valuta le proposizioni in mondi possibili tenendo fisso il mondo attuale come punto di riferimento — e l'operatore **Fixedly** (talvolta denominato **A**, dall'avverbio *actually*) — che valuta le proposizioni in mondi possibili trattati ciascuno come se fosse il mondo attuale. Il risultato è una struttura modale con due «assi» di variazione: la variazione dei mondi possibili come circostanze di valutazione, e la variazione dei mondi possibili come contesti (mondi «attuali» alternativi). Una proposizione è **necessaria a priori** se è vera per ogni combinazione di contesto e circostanza di valutazione. Una proposizione è **necessaria a posteriori** se è vera per ogni circostanza di valutazione tenuto fisso il mondo attuale come contesto, ma non è vera per ogni scelta del mondo come contesto.

Chalmers ha sviluppato la versione filosofica più elaborata del bidimensionalismo, applicata al problema della coscienza e della necessità a posteriori kripkeana. Ogni espressione ha due intensioni: la **intensione primaria** (o **intensione epistemica**) è una funzione dai *mondi centrati* — mondi possibili con un punto di riferimento, un «me qui ora» — ai valori semantici; la **intensione secondaria** (o **intensione metafisica**) è una funzione dai mondi possibili ai valori semantici, tenendo fisso il mondo attuale come punto di «rigidizzazione». Per i nomi propri come «acqua», l'intensione primaria cattura il modo in cui il riferimento viene inizialmente stabilito: «acqua» nel mondo primario si riferisce a ciò che si trova al posto dell'acqua nel contesto del parlante, ovvero a qualunque sostanza abbia le proprietà superficiali dell'acqua nell'esperienza del parlante. L'intensione secondaria, invece, è una funzione rigida: «acqua» denota H_2O in tutti i mondi possibili, perché H_2O è la natura reale della sostanza così come è stata determinata nel mondo attuale.

La verità necessaria a posteriori «acqua è H_2O» emerge dalla discrepanza tra le due intensioni. Primariamente, la proposizione non è necessaria: ci sono mondi centrati in cui la sostanza acquosa è XYZ (come nell'esperimento mentale di Putnam) e la proposizione «acqua è H_2O» sarebbe falsa. Secondariamente, la proposizione è necessaria: dato che nel mondo attuale l'acqua è H_2O, «acqua» rigidamente denota H_2O in ogni mondo possibile. La necessità a posteriori è così il risultato di una proposizione il cui valore di necessità varia a seconda di quale delle due intensioni si consideri: è contingente rispetto all'intensione primaria (l'acqua avrebbe potuto non essere H_2O, se «acqua» si fosse riferita a XYZ) ed è necessaria rispetto all'intensione secondaria (tenuto fisso che l'acqua è H_2O nel mondo attuale, non può essere altrimenti).

La struttura formale del bidimensionalismo può essere esposta attraverso la nozione di **matrice bidimensionale**. Data un'espressione E, si costruisce una matrice $M_E : W \times W \rightarrow V$ dove W è l'insieme dei mondi possibili e V è l'insieme dei valori semantici (valori di verità, individui, proposizioni, ecc.). La prima coordinata rappresenta il mondo come contesto (il mondo «attuale» o il mondo centtrato da cui si parla), la seconda come circostanza di valutazione. $M_E(w_1, w_2)$ è il valore semantico di E quando w_1 è il mondo attuale e si valuta in w_2. L'intensione primaria di E è la diagonale della matrice: $P(E) = \lambda w.\ M_E(w, w)$, il valore di E quando il mondo è sia il contesto che la circostanza. L'intensione secondaria di E rispetto al mondo attuale $w@$ è la sezione verticale: $S(E) = \lambda w.$

M_E(w@, w), il valore di E quando w@ è fisso come contesto e si varia la circostanza. Una proposizione è *a priori* se la sua intensione primaria è costantemente vera (vera in ogni mondo centrato). È *necessaria* se la sua intensione secondaria rispetto al mondo attuale è costantemente vera.

Le critiche al bidimensionalismo sono state formulate con particolare acutezza da Scott Soames e da Francois Recanati. Soames sostiene che il bidimensionalismo, nella versione di Chalmers, equivoca tra la struttura semantica delle espressioni e la struttura epistemica dei concetti che i parlanti associano a tali espressioni: le intensioni primarie non sono il significato delle espressioni ma piuttosto le *descrizioni* che i parlanti associano ai termini per fissarne il riferimento in circostanze ipotetiche, e questo contenuto descrittivo non è il contenuto semantico in senso stretto. Recanati contesta che una singola espressione possa avere due intensioni genuinamente distinte: l'ambiguità apparente tra intensione primaria e secondaria riflette una confusione tra il significato linguistico dell'espressione e l'uso che ne fa il parlante in contesti epistémici particolari. Queste critiche non hanno dissolto il programma bidimensionale ma ne hanno chiarito i limiti e le condizioni di adeguatezza, aprendo un dibattito che resta tra i più vivi nella semantica filosofica contemporanea.

II.5 — Grammatica di Montague: composizionalità, tipi semantici e la formalizzazione del linguaggio naturale

Richard Montague ha realizzato, in una serie di articoli tra il 1970 e il 1973 — «Universal Grammar» (1970), «The Proper Treatment of Quantification in Ordinary English» (PTQ, 1973) — il programma più ambizioso e più tecnicamente elaborato di formalizzazione della semantica del linguaggio naturale. La prima frase di «Universal Grammar» è la dichiarazione di un programma: non vi è alcuna differenza teoricamente significativa tra il linguaggio naturale e i linguaggi formali dei logici e dei matematici, e la semantica di entrambi deve essere trattata allo stesso modo — come interpretazione di un'algebra sintattica astratta attraverso strutture matematiche determinate.

Il sistema montagoniano si fonda su un'algebra ricorsiva dei tipi semantici. A partire da due tipi base — **e** (il tipo delle entità individuali) e **t** (il tipo dei valori di verità {0, 1}) — si generano tutti i tipi complessi attraverso la regola: se σ e τ sono tipi, allora $\langle\sigma, \tau\rangle$ è il tipo delle funzioni totali da oggetti di tipo σ a oggetti di tipo τ. La gerarchia dei tipi è dunque:

e, t sono tipi base; $\langle e, t\rangle$ è il tipo delle funzioni da individui a valori di verità — le proprietà degli individui, o insiemi di individui; $\langle\langle e, t\rangle, t\rangle$ è il tipo delle funzioni da proprietà di individui a valori di verità — le proprietà di secondo livello; $\langle e, \langle e, t\rangle\rangle$ è il tipo delle relazioni binarie tra individui; e così via per ogni combinazione ricorsiva.

La semantica **intensionale** aggiunge una dimensione di mondi possibili e istanti temporali. Un **indice** è una coppia $\langle w, t\rangle$ dove w è un mondo possibile e t un istante temporale. Un'**intensione** di tipo σ è una funzione dagli indici alle estensioni di tipo σ. Il tipo intensionale corrispondente a σ è $\langle s, \sigma\rangle$ dove s è il tipo

degli indici. Il significato pieno (in senso montagoniano) di un'espressione di tipo σ è la sua intensione — una funzione dagli indici alle estensioni.

Le principali categorie del sistema PTQ ricevono le seguenti interpretazioni. I **nomi comuni** (Common Nouns) hanno tipo $\langle e, t\rangle$: «uomo» denota la funzione che manda ogni individuo a 1 se è un uomo, a 0 altrimenti. I **sintagmi verbali** (Verb Phrases) hanno il medesimo tipo $\langle e, t\rangle$: «cammina» denota la funzione che manda ogni individuo a 1 se cammina nell'indice di valutazione. Le **frasi** (Sentences) hanno tipo t: una frase completa denota un valore di verità nell'interpretazione estensionale, o una proposizione (funzione da indici a valori di verità) nell'intensionale.

I **sintagmi nominali** (Noun Phrases) ricevono il tipo $\langle\langle e, t\rangle, t\rangle$ — il tipo dei **quantificatori generalizzati**. Questa è la mossa più originale e più discussa del sistema. Invece di trattare i sintagmi nominali indefiniti («un uomo») e universali («ogni uomo») come quantificatori e i nomi propri («Giovanni») come termini singolari di tipo e, Montague tratta tutti i sintagmi nominali uniformemente come quantificatori generalizzati — funzioni dall'insieme delle proprietà ai valori di verità. Un nome proprio come «Giovanni» viene sollevato al tipo $\langle\langle e, t\rangle, t\rangle$ attraverso l'operazione di **type lifting**:

$[\![\text{Giovanni}]\!] = \lambda P.\ P(\text{giovanni_individuo})$

la funzione che manda ogni proprietà P al valore di verità di P applicata all'individuo giovanni. Le interpretazioni dei principali determinatori sono:

$[\![\text{ogni uomo}]\!] = \lambda P.\ \forall x\ (\text{uomo}(x) \rightarrow P(x))$ $[\![\text{qualche uomo}]\!] = \lambda P.\ \exists x\ (\text{uomo}(x) \wedge P(x))$ $[\![\text{nessun uomo}]\!] = \lambda P.\ \neg \exists x\ (\text{uomo}(x) \wedge P(x))$

Il meccanismo semantico fondamentale è l'**applicazione funzionale**: la denotazione della frase è il risultato dell'applicazione della denotazione del sintagma nominale alla denotazione del sintagma verbale. Per «ogni uomo cammina»:

$[\![\text{ogni uomo cammina}]\!] = [\![\text{ogni uomo}]\!]([\![\text{cammina}]\!]) = (\lambda P.\ \forall x(\text{uomo}(x) \rightarrow P(x)))(\lambda y.\ \text{cammina}(y)) = \forall x(\text{uomo}(x) \rightarrow \text{cammina}(x))$

Per «qualche uomo cammina»:

$[\![\text{qualche uomo cammina}]\!] = [\![\text{qualche uomo}]\!]([\![\text{cammina}]\!]) = (\lambda P.\ \exists x(\text{uomo}(x) \wedge P(x)))(\lambda y.\ \text{cammina}(y)) = \exists x(\text{uomo}(x) \wedge \text{cammina}(x))$

Per «Giovanni cammina»:

$[\![\text{Giovanni cammina}]\!] = [\![\text{Giovanni}]\!]([\![\text{cammina}]\!]) = (\lambda P.\ P(\text{giovanni}))(\lambda y.\ \text{cammina}(y)) = \text{cammina}(\text{giovanni})$

Il trattamento uniforme dei sintagmi nominali permette di derivare le condizioni di verità delle frasi con determinatori diversi attraverso lo stesso meccanismo composizionale — applicazione funzionale — senza trattare diversamente i nomi propri e i quantificatori.

Il **Principio di Composizionalità** — talvolta denominato principio di Frege, sebbene la sua formulazione esplicita sia più tarda — è il fondamento strutturale dell'intero sistema. Esso afferma che il significato di un'espressione complessa è interamente determinato dai significati delle sue parti costituenti e dalla struttura sintattica che le combina. Nel framework algebrico di Montague, questo principio assume la forma di un **omomorfismo** tra l'algebra sintattica e l'algebra semantica. Sia Syn l'algebra delle espressioni linguistiche con le operazioni sintattiche di costruzione (concatenazione, sostituzione, ecc.), e Sem l'algebra delle denotazioni con le operazioni semantiche di composizione (applicazione funzionale, astrazione lambda, ecc.). La funzione di interpretazione $[\![\cdot]\!]$: Syn → Sem deve essere un omomorfismo — deve commutare con tutte le operazioni:

$$[\![\mathrm{Op_syn}(\alpha_1,\ldots,\alpha_n)]\!] = \mathrm{Op_sem}([\![\alpha_1]\!],\ldots,[\![\alpha_n]\!])$$

per ogni operazione sintattica Op_syn e la corrispondente operazione semantica Op_sem. Questo requisito garantisce che la semantica non possa «guardare dentro» la struttura sintattica in modo indiretto: il significato della costruzione è interamente determinato dai significati dei componenti e dall'operazione che li combina.

I contesti intensionali sono trattati attraverso un doppio meccanismo. L'operatore $\cap$ (intensionalizzazione, «up») manda un oggetto di tipo σ all'intensione costante con quel valore — alla funzione dagli indici che restituisce sempre quell'oggetto. L'operatore $\cup$ (estensionalizzazione, «down») valuta un'intensione all'indice corrente. I verbi intensionali come «crede che» richiedono come argomento l'intensione della frase incorporata — non il valore di verità ma la proposizione:

$$[\![\text{crede}]\!] = \lambda P_{\{\langle s,t\rangle\}}.\ \lambda x_e.\ \mathrm{Crede}(x, P)$$

dove P è una variabile di tipo $\langle s, t\rangle$ (una proposizione, ovvero un'intensione di tipo t). L'enunciato «Giovanni crede che l'Astro della Sera sia luminoso» è analizzato come:

$$[\![\text{Giovanni}]\!](\lambda y.\ \mathrm{Crede}(y, \cap[\![\text{l'Astro della Sera è luminoso}]\!]))$$

dove $\cap[\![\text{l'Astro della Sera è luminoso}]\!]$ è l'intensione della frase incorporata — la funzione dagli indici al valore di verità. La sostituzione di «l'Astro del Mattino» per «l'Astro della Sera» produce:

$$[\![\text{Giovanni}]\!](\lambda y.\ \mathrm{Crede}(y, \cap[\![\text{l'Astro del Mattino è luminoso}]\!]))$$

che è una proposizione diversa dalla precedente, poiché le intensioni dei due nomi — pur convergendo nella medesima denotazione nell'indice del mondo attuale — sono funzioni diverse dagli indici ai valori di verità: esistono indici in cui le descrizioni denotano individui diversi, e le intensioni differiscono in corrispondenza di tali indici. Il meccanismo intensionale cattura la non-interscambiabilità nei contesti obliqui che Frege aveva identificato: le intensioni degli enunciati incassati differiscono anche quando le estensioni coincidono nel mondo attuale.

La grammatica montagoniana ha fertilizzato decenni di ricerca in semantica formale. Ma le sue limitazioni documentate — l'anafora interfraseale (la ripresa di

referenti introdotti da indefiniti in enunciati precedenti, che la semantica delle condizioni di verità enunciato per enunciato non può catturare), la quantificazione plurale (le letture collettive contro quelle distributive), la struttura dell'aspetto verbale, i fenomeni di dipendenza dal contesto che trascendono l'indice ⟨mondo, tempo⟩ — hanno motivato una serie di estensioni che hanno integrato e sfidato il programma originale. La Discourse Representation Theory di Kamp (1981), la semantica degli eventi di Davidson (1967) e Parsons (1990), la semantica plurale di Link (1983) e Schein (1993), la semantica dinamica di Groenendijk e Stokhof (1991) — tutte queste teorie condividono l'ispirazione montagoniana di trattare il linguaggio naturale con strumenti formali rigidi, ma divergono su quale sia la struttura formale adeguata e su quale sia il livello al quale deve operare la composizionalità.

II.6 — Teoria dei quantificatori generalizzati: conservatività, monotonicità e gli universali semantici

Il contributo di Jon Barwise e Robin Cooper in «Generalized Quantifiers and Natural Language» (1981) ha sistematizzato e approfondito l'intuizione montagoniana, fornendo un'analisi formale della semantica dei determinatori — le parole come «ogni», «qualche», «nessuno», «la maggior parte», «almeno tre», «esattamente cinque» — che costituisce uno dei risultati più rigorosi della semantica formale novecentesca, con implicazioni tanto per la linguistica quanto per le teorie cognitive dell'acquisizione del linguaggio.

Un **quantificatore generalizzato** Q è una relazione binaria tra insiemi di individui: Q(A)(B) è un valore di verità che dipende dalla relazione tra l'insieme A — denotato dal nome comune che segue il determinatore, la cosiddetta restrizione — e l'insieme B — denotato dal sintagma verbale, il corpo della quantificazione. Le interpretazioni dei principali determinatori della lingua naturale sono:

$[\![\text{ogni}]\!](A)(B) = 1 \leftrightarrow A \subseteq B$ $[\![\text{qualche}]\!](A)(B) = 1 \leftrightarrow A \cap B \neq \varnothing$ $[\![\text{nessuno}]\!](A)(B) = 1 \leftrightarrow A \cap B = \varnothing$ $[\![\text{la maggior parte}]\!](A)(B) = 1 \leftrightarrow |A \cap B| > |A \setminus B|$ $[\![\text{esattamente n}]\!](A)(B) = 1 \leftrightarrow |A \cap B| = n$ $[\![\text{almeno n}]\!](A)(B) = 1 \leftrightarrow |A \cap B| \geq n$ $[\![\text{al più n}]\!](A)(B) = 1 \leftrightarrow |A \cap B| \leq n$

La formulazione dei quantificatori come relazioni binarie tra insiemi — anziché come unari applicati direttamente a predicati, come nella logica del prim'ordine standard — rivela struttura che la logica classica non mette in evidenza.

La proprietà più importante identificata da Barwise e Cooper, proposta come **universale semantico del linguaggio naturale**, è la **conservatività** (conservativity): per ogni determinatore Q dei linguaggi naturali,

$$Q(A)(B) = Q(A)(A \cap B)$$

In altri termini, la valutazione di un enunciato con un quantificatore della lingua naturale dipende soltanto dall'insieme A del nome comune restrittore e dall'intersezione di A con B — non dall'insieme B nella sua totalità. Ciò che è al di fuori del dominio A del nome comune è irrilevante per la valutazione del

quantificatore. «Ogni studente dorme» ha lo stesso valore di verità di «ogni studente [che sia studente] dorme» — la restrizione ridondante non aggiunge nulla. Formalmente, la conservatività segue dal fatto che Q(A)(B) dipende soltanto da come si distribuiscono gli elementi di A tra B e il suo complemento, non da come si distribuiscono gli elementi fuori di A.

La conservatività vale per tutti i determinatori che compaiono nelle lingue naturali studiate finora, e la sua universalità suggerisce che sia un vincolo cognitivo profondo sull'architettura della quantificazione nel pensiero umano. I determinatori non conservativi — come il determinatore immaginario «tutti e soltanto» nel senso di «esattamente coloro che soddisfano sia A che B» — sono agrammaticali nelle lingue naturali, ma la loro grammaticalità o meno non è ovvia dalla sola struttura sintattica: è la semantica che li esclude. Barwise e Cooper propongono che la conservatività sia un universale semantico, e questa proposta è stata verificata e affinata su un ampio campione tipologico di lingue.

La **monotonicità** è la seconda proprietà fondamentale. Un determinatore Q è **monotono crescente destro** (upward entailing right, UE) se:

$$Q(A)(B) = 1 \wedge B \subseteq B' \rightarrow Q(A)(B') = 1$$

Se la proprietà è soddisfatta e si allarga il predicato, la soddisfazione viene preservata. «Qualche studente» è UE: «qualche studente dorme» e «dormire $\subseteq$ riposare» implicano «qualche studente riposa». Un determinatore è **monotono decrescente destro** (downward entailing right, DE) se:

$$Q(A)(B) = 1 \wedge B' \subseteq B \rightarrow Q(A)(B') = 1$$

Se si restringe il predicato, la soddisfazione viene preservata. «Nessuno studente» è DE: «nessuno studente si muove» e «camminare $\subseteq$ muoversi» implicano «nessuno studente cammina». «Ogni studente» è DE nella restrizione e UE nel nucleo: «ogni studente si muove» e «camminare $\subseteq$ muoversi» implicano «ogni studente cammina». «La maggior parte degli studenti» non è monotono in nessuna direzione.

La distinzione tra ambienti DE e UE ha conseguenze dirette per la distribuzione degli **Elementi a Polarità Negativa** (Negative Polarity Items, NPI) — termini come «mai», «affatto», «nemmeno uno», «alcunché» — che in italiano e in molte altre lingue appaiono soltanto in certi contesti sintattici. La generalizzazione di Ladusaw (1979) afferma che gli NPI sono legittimi soltanto in ambienti DE. Questa generalizzazione fa previsioni empiriche verificabili:

«Nessuno studente ha *mai* superato quell'esame» — grammaticale (nessuno crea ambiente DE) «*Qualche studente ha *mai* superato quell'esame» — agrammaticale al di fuori di interrogative (qualche crea ambiente UE) «Se *qualcuno* viene, porta da bere» — grammaticale (l'antecedente di un condizionale è DE)

La generalizzazione di Ladusaw stabilisce così un ponte tra semantica formale (la struttura dell'ordine d'implicazione) e sintassi (la distribuzione degli NPI), mostrando che proprietà semantiche astratte come la monotonicità hanno conseguenze grammaticali osservabili — un risultato paradigmatico per il

programma di semantica formale che cerca di ridurre la grammatica alla struttura logica.

II.7 — Semantica distribuzionale: la geometria del significato e il problema del radicamento

A partire dagli anni Novanta — con accelerazione esponenziale nel secondo decennio del ventunesimo secolo — un approccio strutturalmente diverso alla semantica ha acquisito posizione dominante nella linguistica computazionale e ha posto sfide teoriche acute alla filosofia del linguaggio: la **semantica distribuzionale**. Questo approccio radica il significato non nelle condizioni di verità ma nelle **distribuzioni di contesti di occorrenza** delle espressioni nel discorso. Il significato di un'espressione, in questa prospettiva, è determinato dall'insieme dei contesti in cui essa compare — dalla «compagnia» che frequenta, per usare la formulazione di J. R. Firth (1957).

L'**ipotesi distribuzionale** ha una genealogia che risale al distribuzionalismo di Zellig Harris negli anni Cinquanta: due espressioni semanticamente simili tendono a comparire in contesti simili, e due espressioni che compaiono negli stessi contesti sono semanticamente equivalenti. Questa ipotesi non è meramente pragmatica: essa suggerisce che il significato sia una proprietà relazionale delle espressioni rispetto al corpus — non un legame intrinseco tra un'espressione e il mondo.

L'operazionalizzazione procede attraverso la costruzione di **matrici di co-occorrenza**. Dato un corpus testuale C e un vocabolario $V = \{w_1, ..., w_{|V|}\}$, si costruisce una matrice M di dimensioni $|V| \times |\Gamma|$ dove Γ è l'insieme dei contesti (finestre lessicali, relazioni sintattiche, o documenti). L'elemento $M_{w,c}$ registra la frequenza di co-occorrenza della parola w nel contesto c. La misura di **Pointwise Mutual Information** (PMI) corregge le asimmetrie di frequenza:

$$\mathrm{PMI}(w, c) = \log_2[P(w, c) / (P(w) \cdot P(c))]$$

dove le probabilità sono stimate dal corpus: $P(w, c) = M_{w,c} / N$ con $N = \Sigma_{w',c'} M_{w'c'}$; $P(w) = \Sigma_c M_{wc} / N$; $P(c) = \Sigma_w M_{wc} / N$. Il PMI misura quanto la co-occorrenza osservata di w e c devia dalla co-occorrenza attesa se w e c fossero distribuiti indipendentemente: $\mathrm{PMI} > 0$ indica associazione positiva, $\mathrm{PMI} < 0$ associazione negativa, $\mathrm{PMI} = 0$ indipendenza statistica. La versione **Positive PMI** (PPMI) tronca i valori negativi:

$$\mathrm{PPMI}(w, c) = \max(0, \mathrm{PMI}(w, c))$$

eliminando i segnali statisticamente poco affidabili delle co-occorrenze sub-attese.

La matrice PPMI è densa ma di grande dimensionalità. La riduzione attraverso la **Singular Value Decomposition** (SVD) produce rappresentazioni dense di dimensione $k \ll |\Gamma|$. Data la decomposizione $M \approx U\Sigma V^T$ dove $U \in \mathbb{R}^{|V| \times r}$ ha vettori singolari sinistri, $\Sigma \in \mathbb{R}^{r \times r}$ è diagonale con valori singolari σ_1

$\geq \ldots \geq \sigma_r \geq 0$, e $V \in \mathbb{R}^{|\Gamma| \times r}$ ha vettori singolari destri, la rappresentazione ridotta di rango k è:

$$M_k = U_k \Sigma_k V_k^T$$

dove U_k, Σ_k, V_k indicano le prime k colonne. La rappresentazione vettoriale di ogni parola w è la riga corrispondente in $U_k \Sigma_k$ (o in $U_k \Sigma_k^{1/2}$ in alcune varianti). La **similarità coseno** tra i vettori di due parole $u = [\![w_1]\!]$ e $v = [\![w_2]\!]$ misura la vicinanza semantica:

$$\cos(u, v) = (u \cdot v) / (\|u\| \cdot \|v\|) = (\Sigma_i u_i v_i) / (\sqrt{(\Sigma_i u_i^2)} \cdot \sqrt{(\Sigma_i v_i^2)})$$

La similarità coseno è invariante per scaling scalare: corrisponde al coseno dell'angolo tra i due vettori nello spazio delle rappresentazioni. Valori prossimi a 1 indicano prossimità semantica distribuzionale; valori prossimi a 0 indicano assenza di distribuzione condivisa.

I modelli neurali di embedding — word2vec (Mikolov, Chen, Corrado, Dean, 2013) nella variante Skip-gram — hanno trasformato il campo, rendendo possibile l'apprendimento di rappresentazioni dense di alta qualità su corpora di miliardi di token in tempi computazionali praticabili. Il criterio di addestramento del modello Skip-gram è la massimizzazione della log-verosimiglianza delle parole di contesto dato il termine centrale:

$$J(\theta) = -(1/T) \Sigma_{t=1}^{T} \Sigma_{-m \leq j \leq m, j \neq 0} \log P(w_{t+j} \mid w_t; \theta)$$

dove T è il numero di token nel corpus, m è la dimensione della finestra di contesto, e $\theta = \{v_w, v'_w\}_{w \in V}$ sono i parametri — vettori di embedding di input v_w e di output v'_w. La probabilità condizionale è modellata attraverso softmax:

$$P(w_O \mid w_I) = \exp(v'_{w_O}{}^T \cdot v_{w_I}) / \Sigma_{w=1}^{|V|} \exp(v'_w{}^T \cdot v_{w_I})$$

Per l'efficienza computazionale, si sostituisce il softmax con il **negative sampling**: si campionano k parole «negative» dalla distribuzione $P_n(w) \propto f(w)^{3/4}$ e si addestra il modello a discriminare le parole di contesto reali dai negativi campionati. L'obiettivo per ogni coppia osservata (w_I, w_O) diventa:

$$\log \sigma(v'_{w_O}{}^T v_{w_I}) + \Sigma_{i=1}^{k} \mathbb{E}_{w_i \sim P_n}[\log \sigma(-v'_{w_i}{}^T v_{w_I})]$$

dove $\sigma(x) = 1/(1+e^{-x})$ è la funzione sigmoide. Il modello GloVe (Pennington, Socher, Manning, 2014) ottimizza una fattorizzazione pesata della matrice di log-co-occorrenza:

$$J = \Sigma_{i,j=1}^{|V|} f(X_{ij}) (v_i^T \tilde{v}_j + b_i + \tilde{b}_j - \log X_{ij})^2$$

dove X_{ij} è il conteggio di co-occorrenza di i e j, b_i e $\tilde{b}_j$ sono bias, e f è la funzione di ponderazione:

$$f(x) = (x / x_{max})^{\alpha} \text{ per } x < x_{max} \quad f(x) = 1 \text{ per } x \geq x_{max}$$

con $x_{max} = 100$ e $\alpha = 3/4$ come valori standard. Pennington et al. dimostrano che GloVe è equivalente, in certe condizioni, a una fattorizzazione di rango

ridotto della matrice di PPMI shiftata da log(k) — stabilendo così una connessione teorica tra i due approcci che illumina la struttura comune sottostante.

La questione filosofica più acuta posta dalla semantica distribuzionale è se le rappresentazioni vettoriali catturino il significato o soltanto un'ombra distribuzionale di esso — una proiezione statistica che preserva le relazioni di similarità e di analogia strutturale ma ignora il contenuto verocondizionale e il referente nel mondo. Il problema del **grounding simbolico** — formulato da Stevan Harnad in «The Symbol Grounding Problem» (1990) — è la versione computazionale del problema filosofico: i simboli di un sistema puramente formale non hanno significato intrinseco; il loro significato deve venire «radicato» in qualcosa di non-simbolico — la percezione, l'azione, l'esperienza — perché altrimenti il sistema manipola sintassi senza semantica, come un sistema che conosce il cinese soltanto attraverso le regole per manipolarne i caratteri, senza mai aver avuto contatto con la Cina.

Nel caso della semantica distribuzionale, il problema del grounding è particolarmente acuto: i vettori di parola sono costruiti interamente da relazioni tra parole — il significato di «cane» è determinato dalle sue co-occorrenze con «abbaia», «pelo», «cuccia», «veterinario» — senza mai stabilire una connessione diretta con un cane reale, con l'esperienza di toccare il pelo di un cane, di udire il suo abbaiare. L'intera struttura è chiusa su se stessa: il significato di ogni parola è determinato dalle parole che la circondano, e il significato di quelle parole è determinato dalle parole che circondano quelle, in una circolarità che sembra non toccare mai il mondo. Questa è la critica che Wierzbicka e i sostenitori del Natural Semantic Metalanguage muovono alla semantica distribuzionale: il significato di «dolore» viene «spiegato» attraverso la distribuzione rispetto ad altri termini emotivi, che sono a loro volta definiti distribuzionalmente, in una struttura autoriferita che non si chiude mai sul contenuto esperienziale.

La replica più sofisticata a questa obiezione — che viene elaborata in forme diverse da coloro che sostengono l'adeguatezza delle rappresentazioni distribuzionali per modellare la cognizione — sostiene che il grounding non è assente ma è distribuito nell'intera struttura delle relazioni distribuzionali: il significato non è una proprietà atomica di un'espressione ma una posizione in una rete di relazioni, e questa posizione codifica implicitamente le strutture causali e categoriali del mondo perché il corpus da cui le distribuzioni sono estratte è stato prodotto da parlanti che hanno avuto contatto diretto con il mondo. Il corpus è, in questo senso, un medium che trasmette il grounding: i parlanti che hanno toccato i cani, udito i loro abbai, portato i loro cuccioli dal veterinario, hanno usato le parole in modi che riflettono queste esperienze, e le distribuzioni statistiche catturano il sedimento di queste esperienze.

Questo argomento è parzialmente convincente per il significato lessicale ordinario, ma incontra difficoltà crescenti con il significato di termini astratti — libertà, giustizia, infinito, probabilità — la cui connessione con l'esperienza sensoriale diretta è mediata da strutture concettuali complesse che non si lasciano facilmente ridurre a distribuzione di contesti. Ed è questo il punto in cui la semantica distribuzionale incontra le questioni delle scienze cognitive che si

esamineranno nella Parte Terza: la domanda di come i concetti umani siano radicati nell'esperienza è la stessa domanda di come i simboli linguistici siano radicati nel mondo, posta da un'angolatura cognitiva anziché formale.

La ricerca sulla **geometria dello spazio semantico** ha rivelato strutture di notevole regolarità che vanno al di là della semplice similarità coseno tra vettori di parola. Mikolov e colleghi hanno documentato che lo spazio vettoriale dei word2vec embedding presenta relazioni analogi computabili attraverso operazioni vettoriali: v(«re») - v(«uomo») + v(«donna») ≈ v(«regina»). Questa aritmetica vettoriale del significato — che ha prodotto un'ampia letteratura di ricerca sulle analogie semantiche e strutturali — suggerisce che le strutture concettuali non sono semplicemente codificate come punti in uno spazio ma come relazioni vettoriali tra punti: la struttura di genere, la struttura di intensità, la struttura di temporalità sono codificate come direzioni vettoriali privilegiate nello spazio semantico. Bolukbasi e colleghi (2016) hanno dimostrato che questi sottospazi di direzione possono codificare bias sociali: la direzione «uomo-donna» nello spazio di embedding cattura la struttura di genere ma anche gli stereotipi di genere incorporati nei corpus, e debiasing il vettore di «medico» rispetto a questa direzione riduce l'associazione stereotipata con il genere maschile.

I modelli di embedding **contestualizzati** — BERT, ELMo, e soprattutto i transformer di grandi dimensioni — hanno superato la limitazione fondamentale dei modelli statici come word2vec: ogni parola ha un unico vettore indipendentemente dal contesto, mentre i modelli contestualizzati producono rappresentazioni che variano con il contesto linguistico. La parola «banca» in «deposito in banca» e «la riva della banca del fiume» produce vettori diversi in BERT, anche se ha la stessa forma fonologica: il modello ha appreso a disambiguare i sensi attraverso il contesto. Questo avvicina i modelli contestualizzati alla struttura semantica fregeana — il contesto come determinante della denotazione — e apre la strada a modelli computazionali della polisemia sistematica, del significato dipendente dal contesto e dei meccanismi di adattamento del significato discussi nelle teorie della Semantica Lessicale Generativa di James Pustejovsky.

II.8 — Pragmatica, implicatura e la dimensione cooperativa del significato

La semantica verocondizionale — nella forma della logica intensionale di Montague o nella forma distribuzionale dell'embedding — cattura ciò che un enunciato **dice** in senso stretto: il contenuto proposizionale, le condizioni di verità, la distribuzione dei contesti. Ma la comunicazione umana trasmette sistematicamente contenuti che eccedono il detto letterale: suggerimenti non asseriti, presupposizioni di sfondo, inferenze pragmatiche che il parlante intende comunicare senza esprimerle esplicitamente. La pragmatica si occupa di questa dimensione eccedente del significato — del divario sistematico tra ciò che si dice e ciò che si comunica.

Paul Grice ha fornito, nelle William James Lectures del 1967 (pubblicate nel 1975 con il titolo «Logic and Conversation»), l'analisi più influente di questo divario. Il suo punto di partenza è il **Principio di Cooperazione** (PC): i parlanti in una conversazione si aspettano l'uno dall'altro di contribuire in modo appropriato allo

scopo dello scambio comunicativo. Il PC si articola in quattro famiglie di massime:

la **Massima di Quantità** esige che il contributo sia tanto informativo quanto richiesto, né più né meno; la **Massima di Qualità** esige che il contributo sia veritiero e ben supportato; la **Massima di Relazione** (o Pertinenza) esige che il contributo sia pertinente; la **Massima di Modo** esige che il contributo sia chiaro, non ambiguo, conciso e ordinato.

Un'**implicatura conversazionale** è un contenuto che il parlante trasmette senza asserirlo letteralmente, e che l'ascoltatore inferisce attraverso il ragionamento: «il parlante ha detto X; se rispettasse la massima M, nel contesto C la scelta di dire X implicherebbe Y; non vi è ragione di pensare che stia violando M; dunque il parlante vuole comunicare Y». La struttura dell'inferenza implicaturale è quella di un ragionamento pratico che utilizza la massima come premessa e la razionalità del parlante come assunzione.

L'esempio paradigmatico dell'implicatura scalare — il caso che ha generato la letteratura più ampia — riguarda la scala delle quantità ⟨qualche, molti, la maggior parte, tutti⟩. Se un parlante dice «qualche studente ha superato l'esame», l'ascoltatore infere che non tutti hanno superato — non perché «qualche» significhi «non tutti», ma perché, se il parlante avesse saputo che tutti avevano superato, la Massima di Quantità (essere sufficientemente informativi) lo avrebbe obbligato a dirlo. Il fatto che abbia usato il termine più debole della scala implica, per la razionalità del parlante, che non era in grado di usare il termine più forte. Formalmente, per ogni scala $\langle \alpha_1, ..., \alpha \rangle$ con α_1 meno informativo di α_2 meno informativo di ... meno informativo di α , l'uso di α_i implica conversazionalmente la negazione di α_{i+1}, ..., α :

$$\text{Detto}(\alpha_i) \rightarrow_{impl} \neg \textit{Detto}(a\{i+1\}) \wedge ... \wedge \neg \text{Detto}(\alpha_n)$$

dove →_impl è la relazione di implicatura (non di implicazione logica).

Le implicature conversazionali hanno tre proprietà strutturali che le distinguono dalle implicazioni semantiche. Sono **cancellabili** senza contraddizione: «qualche studente ha superato — anzi, tutti» elimina l'implicatura senza contraddire il detto. Non sono **distaccabili** dalla forma linguistica: la medesima implicatura emerge da qualunque modo di esprimere il medesimo contenuto proposizionale. Sono **calcolabili** razionalmente dal contenuto detto e dal PC, senza postulare un ulteriore livello di significato convenzionale: la loro derivazione è esplicita e ricostruibile.

La cancellabilità distingue le implicature conversazionali dalle **implicature convenzionali** — contenuti che alcune parole trasmettono per convenzione semantica senza far parte delle loro condizioni di verità. Il «ma» in «Pietro è intelligente ma pigro» non contribuisce alle condizioni di verità dell'enunciato (che sono le stesse di «Pietro è intelligente e pigro»), ma veicola convenzionalmente un'implicatura di contrasto tra le due proprietà — un contenuto che non è cancellabile senza violare la convenzione lessicale del «ma».

La **Teoria della Pertinenza** di Dan Sperber e Deirdre Wilson (*Relevance: Communication and Cognition*, 1986; seconda edizione 1995) ha riformulato la pragmatica in termini cognitivi espliciti, riducendo il sistema di massime griciane a un unico principio. Il **Primo Principio della Pertinenza** (cognitivo) afferma che la cognizione umana tende verso la massimizzazione della pertinenza — si organizza per ricavare il massimo di effetti cognitivi con il minimo di sforzo di elaborazione. Il **Secondo Principio della Pertinenza** (comunicativo) afferma che ogni stimolo ostensivo-inferenziale — ogni atto che segnali l'intenzione di comunicare qualcosa — trasmette la presunzione della propria pertinenza ottimale: produrrà i massimi effetti cognitivi per il minimo sforzo dell'ascoltatore, compatibilmente con le capacità del parlante.

La comprensione linguistica è, in questa prospettiva, un processo di **inferenza ottimizzante** che cerca la prima interpretazione dell'enunciato compatibile con le aspettative di pertinenza generate dall'atto comunicativo. L'ascoltatore non decodifica il messaggio seguendo regole convenzionali — come suggerisce implicitamente il modello del codice — ma costruisce attivamente l'interpretazione attraverso la ricerca inferenziale che si arresta al punto in cui le aspettative di pertinenza sono soddisfatte. Questo modello descrive la comprensione come un processo di equilibrio tra effetti cognitivi (gli aggiornamenti del sistema di credenze prodotti dall'interpretazione) e sforzo cognitivo (il costo computazionale della derivazione dell'interpretazione): l'interpretazione più pertinente è quella che massimizza il rapporto effetti/sforzo.

La teoria della pertinenza ha implicazioni dirette — e spesso non esplicitate — per la valutazione dei sistemi artificiali di comprensione del linguaggio. Se la comprensione è ottimizzazione inferenziale guidata da aspettative di pertinenza radicate in una storia cognitiva e in un contesto di azione, allora un sistema che produce output linguisticamente appropriati attraverso l'ottimizzazione statistica di distribuzioni su corpora — come i modelli di linguaggio di larga scala esaminati nella Parte Quarta — è molto diverso da un sistema che comprende nel senso richiedente dalla teoria della pertinenza. La distanza tra i due non è una questione di prestazione ma di architettura: la pertinenza richiede un sistema che valuta gli effetti cognitivi e il costo computazionale rispetto a un modello dello stato cognitivo dell'interlocutore — una teoria della mente dell'altro che i modelli di linguaggio statistico non implementano esplicitamente.

II.9 — Atti linguistici, performatività e la struttura dell'azione comunicativa

La semantica verocondizionale e la pragmatica dell'implicatura condividono la concezione di fondo che il linguaggio serva primariamente a descrivere stati di cose — che la sua funzione principale sia quella di esprimere proposizioni che possono essere vere o false. John Langshaw Austin ha contestato questa concezione descrittivista, mostrando nelle William James Lectures del 1955 (pubblicate postume come *How to do Things with Words*, 1962) che il linguaggio è primariamente uno strumento di azione — che parlare è fare.

Il punto di partenza di Austin è l'osservazione che molti enunciati in prima persona singolare al presente indicativo attivo non descrivono fatti ma li creano nel momento dell'enunciazione: «Ti prometto di venire», «Vi dichiaro marito e moglie», «Giuro di dire la verità», «Nomino questa nave Astro Azzurro». Questi enunciati — che Austin chiama **performativi** — non hanno condizioni di verità nel senso ordinario: non si chiede se siano veri o falsi, ma se riescano o falliscano. Una promessa può essere sincera o insincera, ma non vera o falsa come un'asserzione. Austin contrappone gli enunciati performativi a quelli **constatativi** (descrittivi), che invece esprimono proposizioni suscettibili di verità o falsità.

La distinzione si rivela però instabile: ogni enunciato constativo è anche, in un certo senso, il compimento di un atto — l'atto di asserire. Austin elabora allora una tripartizione più sottile, valida per ogni enunciato linguistico. L'atto **locutorio** è la produzione dell'enunciato con un certo senso e una certa denotazione — il contenuto proposizionale articolato. L'atto **illocutorio** è la forza dell'atto compiuto nel dire: asserire, promettere, interrogare, ordinare, dichiarare, minacciare, congratularsi, scusarsi. L'atto **perlocutorio** è l'effetto causato nell'ascoltatore dalla comprensione dell'atto illocutorio: convincerlo, spaventarlo, rassicurarlo, divertirlo, persuaderlo. La medesima locuzione può avere diverse forze illocutorie («la finestra è aperta» può essere una descrizione, una richiesta di chiuderla, una minaccia di lasciar entrare il freddo); la stessa forza illocutoria può essere realizzata con diverse forme linguistiche (un ordine può essere espresso con un imperativo, un interrogativo retorico, un enunciato del tipo «mi aspetterei che tu...»).

John Searle ha sistematizzato la teoria degli atti illocutori in *Speech Acts* (1969), introducendo la nozione di **condizioni di felicità** (felicity conditions) — le condizioni che devono essere soddisfatte perché un atto illocutorio sia compiuto con successo. Per la promessa, le condizioni di felicità includono: **condizioni preparatorie** (il parlante è in grado di compiere l'azione promessa; l'ascoltatore preferirebbe che la compisse; non è ovvio che la compirà comunque); **condizioni di sincerità** (il parlante intende compiere l'azione); e la **condizione essenziale** (l'enunciato conta come assunzione dell'obbligo di compiere l'azione). Le condizioni di sincerità determinano quale **stato psicologico** il parlante esprime attraverso l'atto: nell'asserzione, una credenza; nella promessa, un'intenzione; nella richiesta, un desiderio; nella scusa, un rammarico.

La classificazione degli atti illocutori in Searle distingue cinque grandi categorie: gli **assertivi** (asserzioni, descrizioni, previsioni — impegnano il parlante alla verità del contenuto); i **direttivi** (domande, ordini, richieste — cercano di indurre l'ascoltatore a compiere un'azione); i **commissivi** (promesse, offerte, impegni — impegnano il parlante a un'azione futura); gli **espressivi** (congratulazioni, scuse, ringraziamenti — esprimono uno stato psicologico del parlante); le **dichiarazioni** (nomina, licenziamento, dichiarazione di guerra — producono il cambiamento che descrivono mediante l'atto stesso del loro proferimento, in un contesto istituzionale appropriato).

La teoria degli atti linguistici rivela un aspetto del significato che né la semantica verocondizionale né la semantica distribuzionale catturano adeguatamente: il significato non è solo ciò che un'espressione rappresenta, ma ciò che fa nel

contesto di un'interazione. La stessa espressione «puoi aprire la finestra?» ha condizioni di verità determinate (è vera se e soltanto se l'ascoltatore è in grado di aprire la finestra), una distribuzione statistica di contesti, e una forza illocutoria — nel contesto appropriato, una richiesta cortese — che non è riducibile né alle condizioni di verità né alla distribuzione. Un sistema che comprende il linguaggio in senso pieno deve rispondere alla forza illocutoria, non soltanto al contenuto proposizionale.

Questo punto ha conseguenze dirette per la valutazione dei sistemi artificiali di elaborazione del linguaggio naturale, che è l'argomento della Parte Quarta. Un sistema che risponde «sì» all'interrogativa «puoi aprire la finestra?» — interpretandola letteralmente come domanda sulla propria capacità — ha frainteso la forza illocutoria dell'atto, anche se ha correttamente analizzato la struttura sintattica e il contenuto proposizionale dell'enunciato. La domanda di se i sistemi artificiali comprendano il linguaggio non è quindi separabile dalla domanda di se essi abbiano accesso alla forza illocutoria degli atti — alla dimensione di azione che il linguaggio possiede e che la semantica da sola non cattura.

Questa domanda costituisce il raccordo tra la Parte Seconda e la Parte Quarta. Ma prima di affrontarla dal lato dei sistemi artificiali, è necessario esaminarla dal lato della mente umana — chiedersi in che modo gli esseri umani acquisiscono, rappresentano e elaborano i contenuti semantici e pragmatici del linguaggio. Questa è la questione delle scienze cognitive, alla quale si volge la Parte Terza.

II.10 — Inferenzialismo semantico: Sellars, Brandom e il significato come ruolo nell'inferenza

La tradizione della semantica referenziale — che da Frege a Russell a Kripke ha collocato il referente al centro della semantica — e la tradizione intensionalista dei mondi possibili — che aggiunge al referente l'intensione come funzione dai contesti ai valori — condividono un'assunzione di fondo che raramente viene messa in questione: che il significato sia primariamente una relazione tra espressioni e ciò che esse rappresentano, e che la competenza semantica sia la capacità di correlare espressioni a loro rappresentata. Il programma **inferenzialista** della semantica — il cui arco va da Wilfrid Sellars a Robert Brandom — rovescia questa priorità con una decisione filosofica radicale: il significato non è dato dalle condizioni di verità o dai referenti, ma è costituito dal ruolo che un'espressione gioca nelle inferenze corrette. Capire un'espressione non significa sapere a cosa si riferisce ma sapere come usarla: quali asserzioni supporta, da quali asserzioni è supportata, quali azioni o conclusioni ne derivano.

Sellars ha articolato le basi di questa posizione in «Empiricism and the Philosophy of Mind» (1956), un testo che Brandom ha giustamente definito «il posto più importante dalla Prima Critica in poi» per la comprensione della conoscenza. La tesi centrale di Sellars è la critica del **Mito del Dato** (*Myth of the Given*): l'assunzione epistemologica secondo cui vi sarebbero esperienze o dati immediati, pre-concettuali, che servono come fondamento delle conoscenze empiriche. Sellars argomenta che non esiste nulla che sia al tempo stesso

epistemicamente fondazionale — capace di fornire giustificazione a inferenze — e che non implichi già stati concettuali con struttura inferenziale. L'esperienza non è il pavimento su cui si costruisce la conoscenza: essa è già articolata concettualmente, già inscritta in relazioni di ragioni, già implicata in una rete di inferenze. Non c'è un momento pre-inferenziale di contatto puro con il dato che preceda la strutturazione concettuale.

La critica del Dato ha conseguenze dirette per la semantica: essa implica che non c'è un livello di significato pre-inferenziale e direttamente costituito dalla relazione con l'esperienza o con il mondo. I termini osservativi — «rosso», «caldo», «duro» — non significano ciò che significano perché denotano proprietà di cui l'esperienza ci dà accesso diretto, ma perché partecipano a pattern inferenziali: «questo è rosso» supporta «questo non è verde» (per la incompatibilità materiale tra rosso e verde), «questo non è trasparente» (per le incompatibilità fisiche), «questo probabilmente riflette circa 700 nm di luce» (per le inferenze empiriche). Il significato di «rosso» è costituito dall'insieme di queste relazioni inferenziali, non da una relazione di denotazione con una proprietà fenomenica.

Robert Brandom ha sistematizzato il programma inferenzialista in *Making It Explicit* (1994), un'opera di portata sistematica paragonabile alle grandi costruzioni della filosofia moderna. Il punto di partenza di Brandom è la distinzione tra **pratiche implicite** e **contenuto esplicito**: la competenza linguistica consiste primariamente in abilità pratiche — la disposizione a inferire, a stabilire impegni e diritti doxastici, a rispondere alle asserzioni degli altri in modi normativamente appropriati — e la semantica esplicita è la tematizzazione riflessiva di queste pratiche. L'asserzione non è la produzione di un suono o di un testo: è l'assunzione di un impegno doxastico rispetto alla comunità discorsiva, l'accettazione di responsabilità inferenziali.

Il nucleo tecnico del sistema di Brandom è il modello del **registro deontico** (*deontic scorekeeping*): ogni partecipante a una pratica discorsiva tiene traccia degli impegni (*commitments*) e dei diritti (*entitlements*) degli interlocutori. Asserire «p» comporta l'assunzione dell'impegno che p — il parlante diventa responsabile di giustificare p, di non asserire q incompatibile con p, di supportare le conclusioni che seguono da p. L'ascoltatore, nel riconoscere l'asserzione, attribuisce l'impegno al parlante e può aggiornare i propri diritti inferenziali: se il parlante è un'autorità affidabile su p, l'ascoltatore acquisisce il diritto di asserire p a propria volta senza giustificazione indipendente. La struttura della comunicazione è essenzialmente normativa: è un sistema di trasferimento e aggiornamento di impegni e diritti all'interno di una comunità discorsiva.

Le **incompatibilità materiali** sono al centro della semantica inferenzialista di Brandom. Due proposizioni sono materialmente incompatibili se l'impegno su una esclude il diritto all'impegno sull'altra: «questo è rosso» e «questo è verde» (nell'intera superficie) sono materialmente incompatibili, non perché siano contraddittorie in senso logico formale, ma perché le pratiche concettuali associate a «rosso» e a «verde» non permettono di applicarle allo stesso oggetto. La negazione logica è definita attraverso la incompatibilità: «non-p» è la proposizione incompatibile con tutte e soltanto le proposizioni incompatibili con

p. La connessione con la struttura formale logica emerge come risultato, non come premessa: la logica è lo strumento che permette di rendere esplicite le relazioni inferenziali implicite nelle pratiche concettuali.

La connessione tra inferenzialismo e semantica distribuzionale — esaminata nella sezione II.7 — è di notevole interesse filosofico. L'ipotesi distribuzionale afferma che il significato di un'espressione è determinato dalla sua distribuzione di contesti di occorrenza; la semantica inferenzialista afferma che il significato è determinato dal ruolo inferenziale. Queste due posizioni non sono opposte ma complementari: i contesti di co-occorrenza nel corpus codificano statisticamente le relazioni inferenziali e di incompatibilità che strutturano il significato. Una parola che co-occorre frequentemente con «ospedale», «medico», «chirurgia» ha un ruolo distribuzionale che riflette le sue relazioni inferenziali con il dominio sanitario. Il corpus è il sedimento delle pratiche inferenziali di una comunità linguistica, e le strutture statistiche che il modello distribuzionale apprende sono — almeno in parte — la traccia computazionale delle strutture normative del registro deontico.

La critica più incisiva all'inferenzialismo è che esso sembra descrivere meglio la competenza linguistica inter-personale — la capacità di partecipare a pratiche discorsive — che il contenuto semantico come struttura rappresentazionale. Paul Horwich e Jason Stanley hanno sostenuto che l'inferenzialismo non riesce a spiegare il **Frege-Geach problem**: se «il dolore è cattivo» esprime un impegno nel contesto assertivo, che cosa esprime in un'occorrenza all'interno di un condizionale come «se il dolore è cattivo, dovremmo evitarlo»? In quest'ultimo contesto non si assume l'impegno che il dolore sia cattivo — si ragiona ipoteticamente. L'inferenzialismo sembra condannato a postulare due significati distinti per la stessa espressione a seconda del contesto sintattico, mentre la semantica verocondizionale attribuisce un'unica condizione di verità che è valutata allo stesso modo in ogni contesto. Brandom ha risposto a questa critica distinguendo tra **impegni robusti** — assunti direttamente — e **impegni subordinati** — generati dall'occorrenza in contesti condizionali — ma la risposta ha incontrato resistenze che mostrano la profondità della tensione tra l'approccio normativo-inferenziale e l'approccio verocondizionale al significato.

II.11 — Semantica Dinamica e Teoria della Rappresentazione del Discorso

La sezione dedicata alla grammatica di Montague si chiude con il riconoscimento esplicito che il programma composizionale formale incontra resistenze strutturali nei fenomeni di anafora interfraseale, nella quantificazione plurale e nelle costruzioni dipendenti dal discorso. Fra le teorie che hanno tentato di rispondere a queste resistenze mantenendo la precisione formale dell'approccio montagoniano, la **Discourse Representation Theory** (DRT), elaborata da Hans Kamp a partire dal 1981 e sviluppata da Kamp e Uwe Reyle nel corso del decennio successivo, costituisce la proposta più sistematicamente ambiziosa e filosofica mente più consequenziale, perché non si limita a rattoppare il programma esistente ma ne rovescia una premessa di fondo: quella secondo cui il significato di un enunciato è una proposizione — un oggetto semantico completamente determinato dall'enunciato stesso, indipendentemente da ciò che

lo precede e lo segue nel discorso. La DRT sostituisce questa premessa con un'immagine radicalmente diversa: il significato è un processo di costruzione incrementale di strutture rappresentazionali che si dispiegano nel tempo del discorso, e il contributo semantico di ogni enunciato è inseparabile dal contesto discorsivo che lo accoglie e che esso, a propria volta, modifica.

Il punto di partenza formale è un'osservazione che risale almeno a Peter Geach e che ha la virtù di essere insieme semplice e intrattabile per la semantica proposizionale standard. Si considerino le seguenti sequenze discorsive:

Un agricoltore è entrato. Egli portava un cappello.

e, nella versione che ha guadagnato il nome di «frase del somaro» (*donkey sentence*):

Ogni agricoltore che possiede un asino lo batte.

Nel primo caso, il pronome «egli» del secondo enunciato si riferisce all'individuo introdotto dall'indefinito «un agricoltore» nel primo. Ma nella semantica montagoniana la denotazione di un sintagma nominale indefinito è un quantificatore generalizzato del tipo $\langle\langle e, t\rangle, t\rangle$, non un termine singolare: «un agricoltore» denota la funzione $\lambda P.\, \exists x(\text{agricoltore}(x) \wedge P(x))$. Una volta che l'enunciato è stato valutato e il suo valore di verità calcolato, non rimane nel sistema semantico nessuna entità a cui il pronome del secondo enunciato possa agganciarsi: il quantificatore esistenziale è «consumato» nella valutazione dell'enunciato in cui compare, e il suo legame con gli enunciati successivi è irrecuperabile nei termini del meccanismo composizionale standard. La frase del somaro aggrava il problema: nella proposizione universale «ogni agricoltore che possiede un asino lo batte», il pronome «lo» si riferisce all'asino introdotto dall'indefinito nella proposizione relativa, ma quest'indefinito è all'interno del sintagma nominale restrittivo del quantificatore universale — una posizione dalla quale, nella logica del primo ordine standard, non è possibile legare variabili all'esterno della proposizione relativa senza produrre forme agrammaticali o semanticamente errate. La traduzione diretta in logica del primo ordine produce: $\forall x(\text{agricoltore}(x) \wedge \exists y(\text{asino}(y) \wedge \text{possiede}(x,y)) \rightarrow \text{batte}(x,y))$, in cui y appare libera nella conseguente — il che è formalmente malformato. La traduzione corretta richiede di portare il quantificatore esistenziale in portata universale: $\forall x \forall y(\text{agricoltore}(x) \wedge \text{asino}(y) \wedge \text{possiede}(x,y) \rightarrow \text{batte}(x,y))$, ma questa formula non si ottiene con nessuna procedura composizionale standard che rispetti la struttura sintattica della frase.

Kamp risponde a questo intreccio di difficoltà non semplicemente aggiustando i tipi semantici o le regole di combinazione, ma costruendo un livello di rappresentazione intermedio tra la struttura sintattica della frase e i modelli verocondizionali. Tale livello è la **struttura di rappresentazione del discorso** (*Discourse Representation Structure*, DRS), un oggetto formale della seguente forma. Una DRS K è una coppia $\langle U_K, Con_K\rangle$ in cui U_K è un insieme finito di **referenti discorsivi** — variabili che rappresentano gli individui introdotti nel discorso — e Con_K è un insieme finito di **condizioni** su quei referenti. Le condizioni possono essere atomiche della forma P(x) o R(x,y) (con P un predicato unario e R binario), oppure complesse. Le condizioni complesse sono di tre tipi fondamentali: la condizione negata $\neg K_1$, ove K_1 è a sua volta una DRS;

la condizione implicativa $K_1 \Rightarrow K_2$, che esprime la struttura delle quantificazioni universali e dei condizionali; e la condizione disgiuntiva $K_1 \vee K_2$. Ogni DRS complessa del tipo condizionale contiene due DRS subordinate, e i referenti discorsivi introdotti nella DRS di antecedente K_1 sono **accessibili** nella DRS di conseguente K_2, nel senso che le condizioni di K_2 possono contenere variabili che compaiono nel dominio di K_1. Questa relazione di accessibilità è il meccanismo fondamentale che la DRT utilizza per trattare l'anafora: un pronome può essere risolto come coreferente a un referente discorsivo introdotto in una DRS accessibile dalla DRS in cui il pronome compare.

La costruzione della DRS per un discorso procede mediante un algoritmo ricorsivo che trasforma la struttura sintattica degli enunciati in aggiornamenti della DRS in costruzione. Per l'enunciato «Un agricoltore è entrato», il referente discorsivo x viene introdotto nel dominio della DRS principale, e le condizioni agricoltore(x) ed entrato(x) vengono aggiunte all'insieme delle condizioni. La DRS parziale al termine di questo enunciato è $\langle \{x\}, \{\text{agricoltore}(x), \text{entrato}(x)\} \rangle$. Quando il secondo enunciato «Egli portava un cappello» viene processato, il pronome «egli» attiva la ricerca di un antecedente nell'insieme dei referenti discorsivi accessibili: x è accessibile perché è stato introdotto nella stessa DRS principale. Si introduce il referente discorsivo y per «un cappello», e si aggiungono le condizioni cappello(y) e portava(x,y). La DRS risultante è $\langle \{x,y\}, \{\text{agricoltore}(x), \text{entrato}(x), \text{cappello}(y), \text{portava}(x,y)\} \rangle$. Il pronome è stato risolto non assegnando un valore al termine singolare in isolamento, ma attraverso il processo discorsivo di identificazione del referente all'interno della DRS in costruzione.

Per la frase del somaro, la struttura sintattica segnala una quantificazione universale che richiede la costruzione di una condizione implicativa $K_1 \Rightarrow K_2$. La DRS di antecedente K_1 è $\langle \{x,y\}, \{\text{agricoltore}(x), \text{asino}(y), \text{possiede}(x,y)\} \rangle$, in cui x e y sono i referenti discorsivi introdotti rispettivamente da «ogni agricoltore» e da «un asino» nella proposizione relativa restrittiva. La DRS di conseguente K_2 è $\langle \{\}, \{\text{batte}(x,y)\} \rangle$, senza referenti propri ma con occorrenze libere di x e y che, per la relazione di accessibilità, vengono interpretate come le variabili introdotte in K_1. La DRS principale è dunque $\langle \{\}, \{K_1 \Rightarrow K_2\} \rangle$. Le condizioni di verità di questa struttura sono determinate nel modo seguente: la condizione implicativa $K_1 \Rightarrow K_2$ è soddisfatta in un modello M da un'assegnazione f se e soltanto se per ogni estensione g di f tale che g soddisfa K_1 in M, esiste un'ulteriore estensione h di g tale che h soddisfa K_2 in M. Poiché le uniche estensioni rilevanti sono quelle che assegnano valori a x e y soddisfacendo le condizioni di K_1, questo equivale alla condizione: per ogni individuo d_1 nel dominio di M che è un agricoltore e per ogni individuo d_2 che è un asino posseduto da d_1, d_1 batte d_2. Il risultato è formalmente equivalente alla traduzione in logica del primo ordine con entrambi i quantificatori universali, ma ottenuto attraverso un meccanismo composizionale che rispetta la struttura sintattica della frase naturale e tratta l'indefinito nella proposizione relativa in modo uniforme con gli indefiniti che introducono referenti discorsivi nelle sequenze anaforiche semplici.

Le condizioni di verità di una DRS K rispetto a un modello $M = \langle D, I \rangle$ sono definite mediante la nozione di **funzione di incassamento** (*embedding function*). Una funzione $f: U_K \rightarrow D$ è un'incassamento di K in M se e soltanto se per ogni

condizione c in Con_K la coppia (M, f) soddisfa c. Le condizioni di soddisfazione sono le seguenti: la condizione atomica P(x) è soddisfatta da (M,f) se f(x) $\in$ I(P); la condizione R(x,y) è soddisfatta se (f(x), f(y)) $\in$ I(R); la condizione $\neg K_1$ è soddisfatta se non esiste g $\supseteq$ f tale che g è un incassamento di K_1 in M; la condizione $K_1 \Rightarrow K_2$ è soddisfatta se per ogni g $\supseteq$ f che è incassamento di K_1 esiste h $\supseteq$ g che è incassamento di K_2; la condizione $K_1 \vee K_2$ è soddisfatta se f è incassamento di K_1 oppure di K_2. La DRS K è **vera** in M se esiste un incassamento di K in M — se la struttura di rappresentazione del discorso è realizzabile nel modello. La scelta di esistere piuttosto che quantificare universalmente sugli incassamenti corrisponde al fatto che i referenti discorsivi introdotti dagli indefiniti sono trattati come variabili esistenzialmente legate dalla condizione di verità della DRS.

Parallelamente alla DRT, Irene Heim ha sviluppato, nella sua dissertazione del 1982, un approccio semanticamente equivalente denominato **File Change Semantics** (FCS). Nella FCS, il contesto semantico è concettualizzato come un «fascicolo» (*file*), un insieme di schede (*file cards*) una per ogni individuo saliente nel discorso, ognuna delle quali porta un insieme di proprietà note di quell'individuo. Un enunciato aggiorna il fascicolo aggiungendo nuove schede per i referenti indefiniti introdotti, o aggiungendo nuove proprietà a schede già esistenti per i referenti resi salienti in precedenza. I pronomi funzionano come descrizioni anaforiché che riattivano schede esistenti. La differenza più significativa tra la DRT e la FCS non è nella struttura formale — i due sistemi sono in larga parte mutuamente traducibili — ma nella metafora computazionale: la DRT modella il discorso come la costruzione incrementale di una struttura logico-formale, la FCS come l'aggiornamento di una base di dati contestuale. Entrambe le metafore catturano aspetti reali della dinamica del significato discorsivo, e la loro complementarità è stata sfruttata in ricerche successive che integrano i due approcci.

La dimensione formale più radicale dell'approccio dinamico è stata sistematizzata da Jeroen Groenendijk e Martin Stokhof nella **semantica dinamica del predicato** (*Dynamic Predicate Logic*, DPL), proposta nel 1991 come alternativa alla DRT che evita il livello intermedio delle DRS e lavora direttamente sulla relazione tra formule e coppie di assegnazioni. Nella DPL, ogni formula φ denota non una proposizione (un insieme di mondi o una funzione dai mondi ai valori di verità) ma una **relazione tra assegnazioni** — l'insieme delle coppie (g, h) in cui g è l'assegnazione di input e h è l'assegnazione di output, modificata dalla formula. Il cuore del sistema è la nuova semantica del quantificatore esistenziale e della congiunzione. Nella DPL:

$$[\![\exists x\, \varphi]\!] = \{(g, h) \mid \exists d: (g[x \mapsto d], h) \in [\![\varphi]\!]\}$$

dove $g[x \mapsto d]$ è l'assegnazione identica a g tranne che per la variabile x, cui viene assegnato il valore d. Il quantificatore esistenziale è dunque un operatore che modifica l'assegnazione di input assegnando un nuovo valore alla variabile x e passa questa assegnazione modificata alla formula nel suo ambito. Per la congiunzione:

$$[\![\varphi \wedge \psi]\!] = \{(g, h) \mid \exists k: (g, k) \in [\![\varphi]\!] \text{ e } (k, h) \in [\![\psi]\!]\}$$

La congiunzione è dunque **composizione sequenziale** di relazioni tra assegnazioni: l'output di φ diventa l'input di ψ. Le formule atomiche agiscono come **test** — relazioni identità condizionali:

$[\![P(x)]\!] = \{(g, g) \mid g(x) \in [\![P]\!]\}$

Non modificano l'assegnazione ma la filtrano: passano l'assegnazione invariata se la condizione è soddisfatta, la bloccano altrimenti. La verità classica di una formula φ rispetto a un'assegnazione g diventa il caso speciale in cui esiste qualche h tale che $(g, h) \in [\![\varphi]\!]$. Questa struttura formale cattura elegantemente la dinamica anaforica: $\exists x(\text{asino}(x)) \wedge \text{grigio}(x)$ è interpretata come la composizione della relazione che introduce una variabile x con valore asinino nell'assegnazione, seguita dal test che filtra le assegnazioni in cui x denota un individuo grigio. La variabile x, una volta introdotta dal quantificatore esistenziale, rimane nell'assegnazione di output e può essere ripresa dalla formula successiva — esattamente come un referente discorsivo introdotto da un indefinito rimane accessibile agli enunciati successivi.

La negazione nella DPL è il punto in cui la dinamicità scontra i limiti dell'accessibilità anaforica. La DPL adotta una semantica classica per la negazione: $[\![\neg\varphi]\!] = \{(g, g) \mid \text{non esiste h: } (g, h) \in [\![\varphi]\!]\}$. La negazione è un test: passa l'assegnazione invariata se φ non è soddisfatta, ma non permette che le variabili introdotte internamente a φ si propaghino all'esterno. Questo è corretto rispetto all'anafora: il discorso «Non ho incontrato un agricoltore» seguito da «*Egli portava un cappello» è agrammaticale nell'uso anaforico, e la DPL lo predice correttamente perché la negazione isola il suo contenuto dalla propagazione dell'assegnazione. Il contrasto con la DRS implicativa è illuminante: nella frase del somaro, il referente discorsivo introdotto nell'antecedente è accessibile nel conseguente perché l'implicazione condizionale della DRT ($K_1 \Longrightarrow K_2$) ha una semantica che permette questa propagazione, mentre la negazione $\neg K_1$ la blocca.

La **semantica degli aggiornamenti** (*Update Semantics*) di Frank Veltman, proposta nel 1996, porta il programma dinamico in una direzione parzialmente diversa: invece di modellare il significato come relazioni tra assegnazioni, lo modella come funzioni da **stati informativi** a stati informativi. Uno stato informativo σ è un insieme di mondi possibili — il contesto epistemico del discorso, l'insieme dei mondi compatibili con l'informazione attualmente disponibile. L'aggiornamento di σ con φ, scritto σ[φ], è il nuovo stato informativo che risulta dall'incorporazione dell'informazione veicolata da φ: $\sigma[\varphi] = \{w \in \sigma \mid w \Vdash_\sigma \varphi\}$, dove la relazione di supporto $\Vdash_\sigma$ è parametrizzata allo stato informativo corrente. Questa parametrizzazione è la novità cruciale: il valore semantico di φ in w dipende non soltanto da w (come nella semantica dei mondi possibili standard) ma dall'intero stato informativo σ in cui la valutazione ha luogo.

L'applicazione più feconda dell'Update Semantics è il trattamento degli **operatori modali epistemici**. Nella semantica dei mondi possibili di Kripke, il modale «potrebbe» (*might*, «forse» in italiano) è analizzato come un quantificatore esistenziale sui mondi accessibili: $\Diamond\varphi$ è vero in w se φ è vera in qualche mondo

accessibile da w. Ma questa analisi ha difficoltà note con i cd. *Moorean paradoxes* dell'asserzione epistemica: un enunciato come «piove, ma forse non piove» risulta contraddittorio in ogni semantica di mondi possibili standard, eppure sembra possibile che un parlante razionale abbia credenze incerte che lo portino ad asserire entrambe le parti in contesti diversi. Nell'Update Semantics, «forse φ» ha la seguente semantica:

$w \Vdash_\sigma \Diamond\varphi$ sse $\sigma[\varphi] \neq \varnothing$

cioè: «forse φ» è supportato dallo stato informativo σ nel mondo w se e soltanto se lo stato informativo σ è compatibile con φ — se aggiornare σ con φ non produce lo stato vuoto (l'assurdo informatico). Crucialmente, la verità dell'enunciato modale non dipende da w ma da σ: è una proprietà dello stato informativo globale, non del mondo di valutazione. Ciò spiega perché «forse piove» è un'affermazione su ciò che il parlante sa (o non sa), non una descrizione oggettiva del mondo: il contenuto del modale è un contenuto di tipo *test* sull'informazione disponibile, non una proposizione sullo stato del mondo.

Questa distinzione tra contenuto proposizionale e contenuto testuale — tra formule che aggiornano lo stato informativo modificandolo e formule che si limitano a testarlo, lasciandolo inalterato se soddisfatto o riducendolo all'insieme vuoto se non soddisfatto — è la base formale dell'analisi veltmaniana della presupposizione. Una presupposizione di un enunciato φ è una condizione P tale che l'aggiornamento $\sigma[\varphi]$ è definito soltanto se P è già soddisfatta dallo stato informativo σ: se P non è soddisfatta, il tentativo di aggiornare produce non lo stato vuoto ma un errore di **presupposizione fallita**. La connessione con la teoria delle presupposizioni di Robert Stalnaker è diretta: il **terreno comune** (*common ground*) di una conversazione è lo stato informativo condiviso dai partecipanti — l'insieme dei mondi compatibili con le proposizioni mutuamente accettate — e ogni atto assertivo presuppone che le sue presupposizioni siano già incorporate nel terreno comune. L'aggiornamento del terreno comune con un'asserzione è l'operazione semantica di base della conversazione, e la struttura formale dell'Update Semantics fornisce il quadro matematico preciso per descrivere questa operazione.

Le implicazioni filosófiche di questo programma dinamico per il problema centrale di questa monografia — il problema di come la struttura raggiunga il contenuto — sono profonde e meritano di essere esplicitate. La semantica proposizionale montagoniana concepisce il significato di un enunciato come una funzione intransitiva: una mappa da indici (coppie mondo-istante) a valori di verità. In questa concezione, il significato è una proprietà dell'enunciato in isolamento, indipendente dal contesto discorsivo, e il contesto influisce sulla semantica soltanto nella misura in cui determina il valore delle espressioni indicali e dei dimostrativi. La semantica dinamica rovescia questa immagine: il significato di un enunciato è la sua **potenzialità di cambiamento del contesto** (*context change potential*) — la funzione che mappa gli stati informativi (o le assegnazioni, o le DRS) pre-enunciazione negli stati informativi post-enunciazione. Il significato non è un oggetto ma un'operazione; non è ciò che l'enunciato descrive ma ciò che l'enunciato fa al contesto discorsivo che lo accoglie.

Questa rivalutazione non è puramente tecnica. Essa riflette una concezione del linguaggio come attività temporalmente distribuita — non come un insieme di atti locali e atomici di descrizione, ma come la costruzione collaborativa e progressiva di strutture condivise di rappresentazione del mondo. La DRS di un discorso non è la somma delle proposizioni degli enunciati che lo compongono: è una struttura integrata in cui le dipendenze anaforiche, le presupposizioni, i contenuti impliciti e i referenti condivisi formano una rete che eccede la semplice giustapposizione delle parti. Il significato discorsivo è, in questo senso, irriducibilmente olista: emerge dalla struttura dell'intero discorso, non dalla somma dei significati delle parti. Questa forma di olismo semantico è distinta dall'olismo delle credenze di Quine — non afferma che il significato di ogni termine dipenda dall'intera rete delle credenze del parlante — ma condivide l'intuizione che la struttura del contesto dia forma al contenuto locale in modi che nessuna semantica puramente composizionale e orientata all'enunciato può catturare.

La connessione con la semantica degli atti linguistici di Austin e Searle, esaminata nella sezione II.9, è qui particolarmente stringente. Se la forza illocutoria di un atto linguistico è ciò che l'atto *fa* nel contesto dell'interazione, e se il significato discorsivo è la potenzialità di cambiamento del contesto, allora la semantica dinamica offre un quadro formale per integrare il contenuto proposizionale (ciò che l'enunciato dice) e la forza illocutoria (ciò che l'atto fa). Un'asserzione aggiorna il terreno comune aggiungendo la proposizione asserita; una domanda apre un'agenda di questioni irrisolte che strutturano le risposte ammissibili; un ordine crea un'obbligazione che aggiorna lo stato normativo della conversazione. La semantica degli aggiornamenti non è soltanto una teoria dell'anafora: è l'abbozzo di una semantica integrata del discorso in cui contenuto proposizionale, forza illocutoria e struttura discorsiva sono aspetti di un unico fenomeno — il significato come trasformazione del contesto.

Il programma inaugurato dalla DRT e sviluppato dalla DPL e dall'Update Semantics è tuttora attivo e produttivo. La **semantica inquisitiva** (*Inquisitive Semantics*) di Groenendijk e Roelofsen, sviluppata nel primo decennio del ventunesimo secolo, estende il quadro dinamico per incorporare la struttura delle domande: una formula denota non una proposizione (un insieme di mondi) ma una questione — un insieme di alternative possibili che la formula apre. Le proposizioni sono casi degeneri di questioni (questioni con un'unica alternativa), e le domande genuine hanno più alternative. La composizionalità si estende alle questioni esattamente come alle proposizioni: la semantica di un enunciato interrogativo si calcola composizionalmente dai significati delle parti, e il significato del testo come sequenza di atti è la serie delle trasformazioni che esso opera sullo stato informativo e sull'insieme delle questioni aperte. Il dialogo diventa così, per questa semantica, il processo di apertura e risoluzione progressiva di questioni — un processo in cui il significato è inseparabile dalla struttura dell'interazione comunicativa che lo genera.

PARTE TERZA: COGNIZIONE, CONCETTI E LA STRUTTURA MENTALE DEL SIGNIFICATO

III.1 — L'intuizionismo di Brouwer: la matematica come costruzione nel tempo interiore e il rifiuto del terzo escluso

La Parte Prima si è conclusa con la domanda di Wigner sull'efficacia irragionevole della matematica, lasciando aperta la questione di che cosa permetta a una struttura formale di raggiungere ciò che le sta al di là. Una delle risposte più radicali a questa domanda non proviene dalla filosofia analitica né dalla fondazione assiomatica, ma dalla tradizione costruttivista che fa capo a Luitzen Egbertus Jan Brouwer e che merita un esame approfondito non soltanto per il suo interesse storico, ma perché il suo nucleo concettuale ricompare, in forme diverse, tanto nella teoria dei tipi di Martin-Löf — esaminata nella sezione I.7 — quanto nelle scienze cognitive contemporanee. L'intuizionismo brouweriano è, in questo senso, il punto d'incontro tra la Parte Prima e la Parte Terza: esso pone il problema del significato matematico come problema della costruzione mentale, e in questo modo anticipa le domande che le scienze cognitive affronteranno su scala più vasta.

Il punto di partenza di Brouwer è una critica alla logica che rovescia la direzione ordinaria del problema ontologico. Per il platonismo e per la teoria assiomatica degli insiemi, la logica classica — con il suo principio del terzo escluso (tertium non datur: $A \vee \neg A$) e il principio di non-contraddizione — è applicabile incondizionatamente all'universo matematico: ogni enunciato matematico è o vero o falso, indipendentemente dal fatto che noi disponiamo o meno di un metodo per stabilire quale dei due casi si verifichi. Brouwer nega questa tesi con una perentorietà che non è contraddetta da alcuna argomentazione tecnica parziale: la logica classica era stata formulata osservando il ragionamento sulle strutture finite, e la sua estensione all'infinito matematico è una generalizzazione ingiustificata, una proiezione di schemi validi in un dominio ristretto su un dominio in cui non si sa se valgano. L'applicazione del principio del terzo escluso all'infinito richiede che si tratti l'infinito come se fosse un dominio dato e completo — come se esistesse un fatto determinato su ogni enunciato riguardante infinitamente molti oggetti — ma questa presupposizione è, per Brouwer, una finzione metafisica non giustificabile.

La matematica autentica, nella concezione brouweriana, non è un insieme di proposizioni che descrivono un dominio di oggetti esistenti indipendentemente, ma un'attività di costruzione che la mente compie nel tempo interiore (*innerlijke tijdbeleving*). Questa locuzione, centrale nelle prime memorie di Brouwer e in particolare nella dissertazione del 1907, denota la struttura temporale della coscienza matematica: il matematico costruisce oggetti matematici in una successione di atti mentali che si dispiegano nel tempo soggettivo, non nel tempo fisico. La struttura fondamentale è la intuizione del due-in-uno (*twee-eenheid*): la percezione simultanea di una distinzione tra un momento precedente e uno successivo, che costituisce il germe di ogni struttura numerica e sequenziale. I

numeri naturali non esistono in un regno platonico; sono costruiti attraverso la ripetizione iterata di questa intuizione della successione.

L'irrazionalità di questa posizione rispetto al platonismo è apparente. Brouwer non sostiene che la matematica sia soggettiva nel senso psicologistico — che i teoremi matematici dipendano dagli stati mentali contingenti di particolari individui. Sostiene che la matematica è la scienza degli oggetti costruiti nel tempo interiore della coscienza, e che questa costruzione è soggetta a vincoli che la rendono intersoggettivamente comunicabile. Il paradosso è questo: la matematica è universale nella sua struttura senza essere platonica nel suo status ontologico, perché l'universalità è garantita dall'identità della struttura dell'intuizione temporale tra soggetti diversi — una tesi che, ripresa dal kantismo, si espone alle obiezioni standard sul carattere contingente della struttura cognitiva umana, ma che Brouwer sostiene senza tentare di fondarla empiricamente.

La conseguenza tecnica più radicale dell'intuizionismo è il rifiuto del principio del terzo escluso come legge logica universale. Per Brouwer, $A \vee \neg A$ non è valido in generale perché la verità di A richiede una dimostrazione costruttiva di A — una procedura effettiva che produca un oggetto matematico del tipo appropriato — e la verità di $\neg A$ richiede una dimostrazione costruttiva dell'impossibilità di dimostrare A, ovvero una procedura che trasformi qualsiasi (ipotetica) dimostrazione di A in una contraddizione. Non vi è alcun principio logico che garantisca che per ogni A uno dei due sia il caso: potremmo trovarci in una posizione in cui non abbiamo né una dimostrazione di A né una dimostrazione dell'impossibilità di A, e in tale posizione il terzo escluso semplicemente non si applica.

L'interpretazione di Brouwer-Heyting-Kolmogorov (BHK) formalizza questa idea in termini di ciò che costituisce una dimostrazione costruttiva di un enunciato composto. La sua struttura è precisamente la corrispondenza di Curry-Howard che si è esaminata nella sezione I.7, vista però dal lato filosofico anziché da quello formale. Una dimostrazione di $A \wedge B$ è una coppia ordinata (p, q) dove p è una dimostrazione di A e q è una dimostrazione di B; una dimostrazione di $A \vee B$ è una coppia $(0, p)$ con p dimostrazione di A oppure una coppia $(1, q)$ con q dimostrazione di B — il che garantisce che per dimostrare una disgiunzione si debba sapere quale disjunct è vero. Una dimostrazione di $A \to B$ è una procedura effettiva — una funzione matematicamente precisa — che trasforma ogni dimostrazione di A in una dimostrazione di B. Una dimostrazione di $\bot$ (l'assurdo) non esiste. La negazione $\neg A$ è abbreviazione di $A \to \bot$. Nessuna clausola dell'interpretazione BHK permette di dimostrare $A \vee \neg A$ per un A arbitrario: per farlo bisognerebbe saper scegliere tra p (dimostrazione di A) e q (procedura che trasforma dimostrazioni di A in dimostrazioni di $\bot$), il che equivale a disporre della dimostrazione di uno dei due, che è esattamente ciò che non si ha in generale.

La struttura formale della logica intuizionista che cattura queste condizioni è il calcolo dei sequenti di Gentzen LJ, ottenuto dal calcolo classico LK eliminando la regola di indebolimento a destra nelle regole che introducono la negazione. Il risultato è che le regole per la negazione e la disgiunzione sono asimmetriche rispetto al sistema classico. La regola di eliminazione del doppio negativo — dalla

doppia negazione $\neg\neg A$ si può derivare A in logica classica — non è derivabile in LJ. Si ha invece il più debole risultato che $A \to \neg\neg A$ è derivabile in LJ (ossia, una dimostrazione di A fornisce una procedura che trasforma qualsiasi procedura che trasforma dimostrazioni di A in assurdo essa stessa in assurdo), ma non il converso. La relazione tra logica classica e logica intuizionista è catturata dalla traduzione di Gödel-Gentzen (1933): ogni formula classica A è soddisfatta nella logica classica se e soltanto se la sua traduzione A^N — ottenuta premettendo una doppia negazione a ogni sottoformula atomica — è derivabile nella logica intuizionista. In notazione:

$$\mathrm{CL} \vdash A \iff \mathrm{IL} \vdash A^N$$

dove A^N è la formula con doppia negazione di ogni atomica: $p^N = \neg\neg p$, $(A \wedge B)^N = A^N \wedge B^N$, $(A \vee B)^N = \neg(\neg A^N \wedge \neg B^N)$, $(A \to B)^N = A^N \to B^N$. Questo risultato mostra che la logica classica è contenuta nella logica intuizionista nel senso tecnico preciso che la traduzione fornisce: non si perde dimostrative power, ma si cambia la natura delle dimostrazioni.

La semantica dei modelli di Kripke per la logica intuizionista — distinta e più sottile rispetto alla semantica modale esaminata nella sezione II.3 — cattura formalmente il carattere temporale dell'intuizionismo brouweriano. Un frame intuizionistico di Kripke è una coppia $\langle W, \leq \rangle$ dove W è un insieme di «stati di conoscenza» (o «mondi possibili informazionali») e $\leq$ è un ordine parziale su W che rappresenta l'avanzamento della conoscenza nel tempo. La condizione di monotonia afferma che la verità si preserva all'avanzare della conoscenza: se $w \leq v$ e $w \vDash p$ (p atomico), allora $v \vDash p$. Le clausole semantiche per i connettivi sono:

$w \vDash A \wedge B \leftrightarrow w \vDash A$ e $w \vDash B$
$w \vDash A \vee B \leftrightarrow w \vDash A$ o $w \vDash B$
$w \vDash A \to B \leftrightarrow \forall v \geq w$: $v \vDash A$ implica $v \vDash B$
$w \vDash \neg A \leftrightarrow \forall v \geq w$: $v \nvDash A$

Il terzo escluso $A \vee \neg A$ non è valido in questa semantica: si può costruire un frame con due stati $w < v$ in cui $v \vDash A$ ma $w \nvDash A$, e allora $w \nvDash A$ (la verità di A non è ancora stabilita in w) e $w \nvDash \neg A$ (la falsità di A non è garantita per tutti i successori di w, perché $v \vDash A$), dunque $w \nvDash A \vee \neg A$. I modelli di Kripke intuizionistici sono dunque frames in cui l'informazione cresce monotonamente e la decidibilità di ogni enunciato non è garantita a priori.

Il teorema del bar (*bar theorem*) di Brouwer è il risultato tecnico più controverso e più profondo dell'intuizionismo, e merita menzione perché illustra come il costruttivismo giunga a conseguenze incompatibili con la matematica classica su punti specifici della teoria delle sequenze infinite. Una *bar* è un insieme B di sequenze finite di numeri naturali tale che ogni sequenza infinita ha almeno un segmento iniziale in B. Il teorema del bar afferma che se B è una bar *decidibile* (per ogni sequenza finita si può decidere se appartiene a B), allora B è una bar *uniforme* — esiste un N tale che ogni sequenza finita di lunghezza N appartiene a B. Nella matematica classica, questo equivale alla compattezza dello spazio di Baire — ogni ricoprimento aperto dello spazio delle sequenze infinite ha un sottocoprimento finito — che è equivalente all'assioma della scelta per sequenze numerabili. Nell'intuizionismo, il teorema del bar è accettato come principio

autonomo e ha la conseguenza che ogni funzione totale definita intuizionisticamente sui reali è uniformemente continua su intervalli compatti — un risultato che contraddice la matematica classica, la quale ammette funzioni totali non continue su $\mathbb{R}$.

Queste conseguenze mostrano che l'intuizionismo non è semplicemente un'aritmetica indebolita rispetto a quella classica ma un universo matematico diverso, con strutture e teoremi differenti. La domanda filosofica cruciale — se l'intuizionismo corrisponda meglio alla struttura effettiva della cognizione matematica umana rispetto al platonismo — è la domanda che connette questa sezione a quelle seguenti: se la matematica è costruzione mentale, la struttura di tale costruzione deve essere studiata empiricamente nelle scienze cognitive, non dedotta a priori da principi astratti.

III.2 — La geometria dei concetti: prototipicità, spazi concettuali e vincoli di convessità

La psicologia cognitiva degli anni Settanta ha prodotto, attraverso il programma sperimentale di Eleanor Rosch e dei suoi collaboratori, una revisione radicale del modello tradizionale della categorizzazione che la filosofia analitica aveva ereditato da Aristotele e mantenuto quasi intatto attraverso la semantica verocondizionale. Il modello aristotelico della categoria — che Wittgenstein aveva già messo in discussione con il concetto di «somiglianza di famiglia» nelle *Ricerche filosofiche* — è il modello delle condizioni necessarie e sufficienti: un oggetto appartiene a una categoria se e soltanto se possiede tutte le proprietà che definiscono la categoria. Un uccello è un uccello se e soltanto se ha piume, è vertebrato, ha un becco, ha due ali, è oviparo: queste condizioni sono la definizione della categoria, e la loro soddisfazione è questione binaria, senza gradazioni.

Rosch dimostrò sperimentalmente, attraverso una serie di studi pubblicati tra il 1973 e il 1978, che le categorie semantiche naturali non hanno questa struttura. Le categorie hanno invece una struttura radiale o prototipica: alcuni membri sono giudicati più tipici di altri, e questa tipicità non è casuale o idiosincratica ma stabile tra soggetti diversi e predittiva di una serie di effetti di elaborazione cognitiva. Il pettirosso è giudicato come uccello più tipico del pinguino, sebbene entrambi soddisfino le condizioni biologiche necessarie per essere uccelli; la mela è un frutto più tipico dell'oliva; una sedia è un mobile più tipico di un candelabro. Il prototipo della categoria è l'esemplare che massimizza la somiglianza familiare con gli altri membri della categoria e minimizza la somiglianza con i membri delle categorie contrastanti.

La somiglianza familiare è la nozione wittgensteiniana formalizzata quantitativamente: per ogni membro m della categoria C e per ogni proprietà saliente f, si assegna un punteggio $s(m, f)$ proporzionale al numero di altri membri che condividono f. La somiglianza familiare di m è la somma dei punteggi su tutte le proprietà salienti:

$$\mathrm{SF}(m) = \Sigma_{f} \left[s(m, f) \cdot n(f, C) \right]$$

dove n(f, C) è la frequenza della proprietà f nella categoria C. Il membro con la massima somiglianza familiare è il prototipo. Il pinguino ha bassa somiglianza familiare rispetto agli altri uccelli perché non condivide con loro proprietà salienti come il volo, il nido sugli alberi, le piccole dimensioni: ne condivide soltanto alcune (piume, becco, oviparo), le meno diagnostiche. I giudizi di tipicità correlano con la somiglianza familiare, con i tempi di risposta in compiti di categorizzazione, con la facilità di apprendimento delle categorie, con l'ordine di acquisizione nella prima infanzia.

Rosch distingue i livelli di astrazione della categorizzazione: il livello di base (*basic level*) è quello in cui le categorie massimizzano la somiglianza intra-categoriale e minimizzano la sovrapposizione inter-categoriale. «Cane» è al livello di base; «animale» è al livello superordinato, troppo astratto; «labrador» è al livello subordinato, troppo specifico. Il livello di base è cognitivamente privilegiato: è il livello al quale si risponde più velocemente in compiti di categorizzazione, si acquisiscono prima i termini lessicali, si producono spontaneamente le denominazioni in assenza di contesto che forzi i livelli adiacenti.

Peter Gärdenfors ha fornito, in *Conceptual Spaces: The Geometry of Thought* (2000), la formalizzazione matematica più coerente di questi risultati, proponendo una teoria dei concetti che occupa una posizione intermedia tra il livello simbolico (la semantica formale proposizionale) e il livello sub-simbolico (le reti neurali connessioniste). Uno **spazio concettuale** è un insieme di dimensioni qualitative Q_1, Q_2, ..., Q_n dotate di struttura metrica o topologica, che rappresentano le dimensioni salienti lungo le quali gli oggetti si distribuiscono. Le dimensioni qualitative corrispondono a proprietà percettive o concettuali: le coordinate di colore nel sistema CIE (luminanza L*, crominanze a*, b*) formano uno spazio tricromatico; le dimensioni di temperatura, durezza, peso, turgidità formano uno spazio per la frutta; le dimensioni di altezza, peso, età formano uno spazio per le persone. Lo spazio concettuale è lo spazio metrico prodotto $\langle Q_1 \times Q_2 \times ... \times Q_n, d \rangle$ dove d è una metrica (tipicamente euclidea o di Minkowski) che integra le dimensioni.

Un **concetto naturale** è, nella proposta di Gärdenfors, una **regione convessa** dello spazio concettuale. Formalmente, per uno spazio metrico $\langle X, d \rangle$, un insieme $C \subseteq X$ è convesso se per ogni $x, y \in C$ e ogni $t \in [0, 1]$:

$$(1-t) \cdot x + t \cdot y \in C$$

dove $(1-t) \cdot x + t \cdot y$ denota il punto sulla geodetica tra x e y che divide il segmento nella proporzione t:(1-t). L'ipotesi di convessità di Gärdenfors afferma che tutte le categorie naturali di oggetti — le categorie esprimibili da singoli aggettivi e sostantivi delle lingue naturali — corrispondono a regioni convesse nei loro spazi concettuali pertinenti. La convessità è un vincolo cognitivo: se un oggetto x e un oggetto y appartengono entrambi alla categoria C, allora ogni oggetto z che si trova «tra» x e y — che è simile a entrambi in tutte le dimensioni rilevanti nella proporzione appropriata — appartiene anch'esso a C.

Il **prototipo** di una categoria convessa C nello spazio concettuale è il centro geometrico (centroide) della regione:

$$\text{proto}(C) = (1/|C|) \int_C x \, d\mu(x)$$

per una misura μ sulla regione C. Il centroide massimizza la somma pesata delle similarità con tutti i punti della regione, e quindi corrisponde operativamente alla nozione intuitiva di prototipo come esemplare che assomiglia massimamente all'insieme dei membri della categoria. Il **grado di appartenenza** di un oggetto a alla categoria C è una funzione decrescente della distanza di a dal centroide proto(C):

$$\mu_C(a) = f(d(a, \text{proto}(C)))$$

dove f è una funzione decrescente (tipicamente esponenziale negativa o lineare troncata). Questo formalizza la struttura prototipica: l'appartenenza non è binaria ma graduata, e il grado è proporzionale alla prossimità al centro della regione.

La convessità ha un'implicazione directa sulla struttura degli spazi semantici costruiti mediante embedding distribuzionale. Se le regioni semantiche nei spazi vettoriali ad alta dimensionalità corrispondono a categorie naturali, e se i concetti naturali sono convessi, allora le regioni semantiche nei spazi di embedding dovrebbero essere convesse rispetto alla metrica coseno. Questa ipotesi — che costituisce una connessione diretta tra la semantica distribuzionale esaminata nella sezione II.7 e la teoria dei concetti di Gärdenfors — è verificata parzialmente nei modelli di embedding moderni: le interpolazioni lineari tra vettori di parola tendono a produrre vettori semanticamente coerenti con le categorie dei vettori interpolati. Il punto medio tra il vettore di «cane» e quello di «lupo» cade in prossimità di «volpe» — un membro intermedio della categoria dei canidi — il che suggerisce convessità locale nello spazio semantico distribuzionale.

La teoria di Gärdenfors consente di dare conto della struttura composizionale del significato senza ridurla alla logica proposizionale. Un sintagma aggettivo-sostantivo come «fragola piccola» denota la regione dell'intersezione tra la regione delle fragole nello spazio concettuale rilevante e la regione degli oggetti piccoli rispetto alla scala delle fragole (non rispetto alla scala degli oggetti in generale). Il concetto di «piccolo» non ha un'estensione assoluta ma è relativo alla categoria: piccolo per un elefante è grande per un topo. Nello spazio concettuale, questo si formalizza come un'operazione di restrizione della regione pertinente alla sotto-regione che soddisfa simultaneamente i vincoli di entrambe le proprietà.

III.3 — La cognizione incarnata: schemi d'immagine, metafore primarie e la radice corporea del significato astratto

La teoria dei concetti di Gärdenfors lascia aperta una questione fondamentale: da dove provengono le dimensioni qualitative degli spazi concettuali? La risposta che Gärdenfors tende a dare — che esse riflettono la struttura dello spazio percettivo, con le sue dimensioni di colore, forma, dimensione, movimento — rinvia a un problema più profondo, che George Lakoff e Mark Johnson hanno formulato in modo esplicito e sistematico in *Metaphors We Live By* (1980) e in *Philosophy in the Flesh* (1999): il significato — e in particolare il significato dei concetti astratti — è radicato nell'esperienza corporea.

La tesi centrale della semantica cognitiva di Lakoff e Johnson, denominata «embodied realism», è che la cognizione non è il prodotto di un processore astratto che manipola simboli formali privi di contenuto intrinseco, ma è strutturata da **schemi d'immagine** (*image schemas*) — strutture pre-concettuali ricorrenti che emergono dall'esperienza sensomotoria e servono come impalcatura per la concettualizzazione di domini astratti. Gli schemi d'immagine non sono immagini nel senso di rappresentazioni visive particolari, ma pattern strutturali generali che si manifestano attraverso l'interazione corporea con l'ambiente: il CONTENITORE (essere dentro o fuori qualcosa), il PERCORSO (movimento da un punto di partenza a un punto di arrivo attraverso un tragitto), la FORZA (essere spinto, tirato, bloccato), la VERTICALITÀ (su-giù), il COLLEGAMENTO (connessione tra elementi), il CICLO (struttura chiusa che ritorna al punto di partenza).

Questi schemi emergono dallo sviluppo sensomotorio: il bambino sperimenta ripetutamente l'ingresso e l'uscita da contenitori (il bagnetto, le braccia della madre, la culla), il percorso verso un oggetto desiderato, la resistenza di oggetti che non si spostano con la forza esercitata. Questa esperienza ripetuta solidifica strutture cognitive che diventano disponibili come «pattern» astratti applicabili a domini diversi dall'esperienza corporea originaria attraverso la **metafora concettuale**. Una metafora concettuale non è una figura retorica linguistica ma una mappatura sistematica tra un dominio sorgente — strutturato dall'esperienza sensoriale — e un dominio bersaglio — tipicamente più astratto. La mappatura preserva la struttura inferenziale del dominio sorgente: le inferenze valide nel dominio sorgente diventano, attraverso la metafora, inferenze disponibili nel dominio bersaglio.

L'esempio paradigmatico è IL TEMPO È LO SPAZIO. Questa metafora concettuale non è un'aggiunta ornamentale al linguaggio temporale ma la sua struttura portante: parliamo del futuro come di qualcosa davanti a noi («guardare al futuro», «il futuro è alle porte»), del passato come di qualcosa dietro di noi («voltarsi al passato»), della durata come di un'estensione spaziale («un lungo periodo», «un breve intervallo»), dell'approssimarsi di eventi come del movimento di oggetti verso il punto di osservazione («si avvicina il Natale», «la scadenza è imminente»). La mappatura proietta la struttura spaziale — le relazioni di vicinanza, distanza, direzione, dimensione — sul dominio temporale, e le inferenze spaziali diventano inferenze temporali disponibili: se qualcosa è lontano nello spazio richiede tempo per raggiungerlo, analogamente qualcosa di lontano nel tempo richiede attesa per raggiungerlo.

La struttura formale della metafora concettuale è una famiglia di corrispondenze ontologiche e corrispondenze epistemiche. Sia S il dominio sorgente e T il dominio bersaglio. La metafora M è una funzione parziale $M: S \to T$ che associa entità, relazioni e schemi inferenziali di S a entità, relazioni e schemi inferenziali di T. Per la metafora IL TEMPO È LO SPAZIO:

M(posizione_spaziale) = istante_temporale
M(distanza_spaziale) = durata_temporale
M(direzione_davanti) = futuro

M(direzione_dietro) = passato
M(velocità_di_moto) = velocità_del_tempo

Le inferenze dello schema PERCORSO nello spazio sorgente — se si percorre un tragitto in linea retta, ciò che si raggiunge prima è più vicino al punto di partenza — diventano inferenze temporali disponibili: gli eventi che si verificano prima nel tempo sono più prossimi al presente. La mappatura proietta la topologia dello schema PERCORSO sulla struttura del dominio temporale.

La teoria delle **metafore primarie** di Joseph Grady (1997) raffina ulteriormente l'analisi, distinguendo tra le metafore concettuali semplici, radicate in correlazioni dirette nell'esperienza incarnata, e le metafore complesse, costruite per composizione delle primarie. Le metafore primarie originano dalla co-occorrenza sistematica di esperienze soggettive (*subjective judgments*) ed esperienze sensomotorie (*sensorimotor experiences*) durante lo sviluppo. La metafora primaria L'AFFETTO È CALORE è fondata sulla co-occorrenza, nell'esperienza infantile, della percezione termica del calore corporeo della madre e dell'esperienza soggettiva dell'affetto. La metafora primaria LE CATEGORIE SONO CONTENITORI è fondata sulla co-occorrenza tra l'esperienza di essere dentro o fuori contenitori fisici e l'esperienza cognitiva di inclusione o esclusione categoriale. Queste correlazioni nell'esperienza producono, attraverso l'apprendimento associativo, un sistema di mappature metaforiche che struttura il pensiero astratto.

Questo programma è stato radicalizzato da Francisco Varela, Evan Thompson e Eleanor Rosch in *The Embodied Mind* (1991), che propone il paradigma dell'**enattivismo**: la cognizione non è la computazione di rappresentazioni interne sulla base di input sensoriali, ma il prodotto dell'accoppiamento sensomotorio (*sensorimotor coupling*) tra organismo e ambiente. Il significato non è una relazione tra una rappresentazione interna e un oggetto esterno ma una proprietà dell'interazione tra organismo e ambiente: emerge dalla storia del loro accoppiamento. L'enattivismo si distanzia tanto dal cognitivismo computazionale (che concepisce la mente come processore di rappresentazioni simboliche) quanto dal connessionismo ingenuo (che sostituisce le rappresentazioni simboliche con pattern di attivazione ma mantiene la struttura input-elaborazione-output), proponendo invece che la mente sia un sistema sensomotorio che genera il proprio mondo di significati attraverso l'azione.

Le implicazioni per la semantica sono dirette. Se il significato emerge dall'accoppiamento sensomotorio, allora la semantica di un termine come «afferrare» non può essere catturata né dalla sua posizione in uno spazio distribuzionale né dalla sua condizione di verità verofunzionale: richiede di essere connessa alla struttura sensomotoria dell'azione di afferrare — alla cinestesi muscolare, alla calibrazione visuo-motoria, alla sequenza di adattamento della presa alla forma dell'oggetto. La semantica distribuzionale che si costruisce su corpora testuali non può, in linea di principio, catturare questo strato del significato, perché i corpora testuali non contengono informazioni sensomotorie. Questo è un versione del problema del grounding che si è esaminato nella sezione II.7, vista ora dalla prospettiva cognitiva anziché da quella formale.

III.4 — Relatività linguistica: la struttura della lingua come vincolo cognitivo

La domanda se il linguaggio modelli il pensiero o il pensiero generi il linguaggio ha attraversato la linguistica e la psicologia cognitiva del Novecento con la persistenza di un problema che non si lascia risolvere facilmente perché i due fenomeni sono così strettamente intrecciati che separarli sperimentalmente richiede una creatività metodologica considerevole. L'ipotesi di relatività linguistica — che prende il nome da Edward Sapir e dal suo allievo Benjamin Lee Whorf, sebbene le sue formulazioni siano precedenti e successive a entrambi — afferma che la struttura della lingua in cui si parla influenza sistematicamente processi cognitivi come la percezione, la memoria, il ragionamento e l'attenzione.

La versione forte dell'ipotesi, il **determinismo linguistico**, afferma che la lingua determina il pensiero nel senso che concetti non lessicalizzati nella propria lingua sono inaccessibili al pensiero. Questa versione è stata abbandonata come insostenibile: i parlanti possono pensare concetti che la loro lingua non lessicalizza, imparare nuove lingue, tradurre tra lingue strutturalmente molto diverse, comunicare su contenuti che nessuna lingua ha codificato prima. La versione debole, la **relatività linguistica**, afferma invece che le strutture della lingua influenzano alcuni processi cognitivi in modi statisticamente misurabili, senza necessariamente escludere l'accesso a concetti non lessicalizzati. È questa versione che ha ricevuto negli ultimi decenni un sostegno sperimentale considerevole, sebbene controverso.

Il dominio del colore ha fornito il banco di prova più elaborato per la relatività linguistica, proprio perché la struttura fisica dello spazio dei colori — determinata dall'anatomia dei coni retinici e dalla neurobiologia del sistema visivo — è uguale per tutti gli esseri umani, il che permette di controllare le variabili fisiche isolando l'effetto della lingua. Lo spazio dei colori del CIE — definito dalle coordinate L*, a*, b* dove L* misura la luminanza e a*, b* misurano le componenti di crominanza — è uno spazio metrico continuo in cui piccole distanze corrispondono a differenze di colore appena percepibili (JND, *just noticeable differences*). Le categorie cromatiche delle lingue suddividono questo spazio continuo in regioni, ma le suddivisioni variano considerevolmente tra le lingue.

L'esperimento di Winawer e colleghi (2007) sfrutta una caratteristica della lingua russa che la distingue dall'inglese: il russo possiede due termini cromatici di base per il blu — «goluboy» (blu chiaro) e «siniy» (blu scuro) — laddove l'inglese ha un unico termine «blue» che copre l'intero spettro. Se la categorizzazione linguistica influenza la discriminazione percettiva, i parlanti russi dovrebbero mostrare un effetto di accelerazione nella discriminazione di due colori che cadono a cavallo del confine goluboy/siniy rispetto a due colori egualmente distanti ma appartenenti alla stessa categoria cromatica. I risultati confermano questa previsione: i parlanti russi sono significativamente più veloci nella discriminazione di colori inter-categoriali rispetto a quelli intra-categoriali, con una differenza di circa 124 millisecondi nei tempi di risposta. I parlanti inglesi non mostrano questo effetto categoriale nella stessa regione dello spazio cromatico. Crucialmente, l'interferenza verbale — un compito verbale concorrente che

occupa le risorse linguistiche del partecipante — elimina il vantaggio nella discriminazione inter-categoriale dei parlanti russi, mentre l'interferenza spaziale non-verbale non lo elimina. Questo dimostra che l'effetto categoriale è mediato da risorse linguistiche, non da differenze percettive di basso livello: è la lingua che accelera la discriminazione, non una diversa sensitività sensoriale.

Il dominio della codificazione spaziale offre un esempio ancora più radicale di variazione inter-linguistica con effetti cognitivi documentati. I sistemi di riferimento spaziale — le cornici di riferimento usate per localizzare oggetti nello spazio — variano tra le lingue in modo sistematico. Il sistema di riferimento **egocentrico** o **relativo** (prevalente nelle lingue europee come l'italiano e l'inglese) descrive la posizione degli oggetti in relazione all'orientamento del parlante: «la palla è a sinistra del tavolo» significa a sinistra rispetto all'orientamento corrente del parlante. Il sistema di riferimento **assoluto** o **allocentrico** (prevalente in lingue come il Guugu Yimithirr australiano o il Tzeltal maya) descrive la posizione degli oggetti in relazione a direzioni fisse dello spazio, tipicamente i punti cardinali: equivalente a «la palla è a nord del tavolo» indipendentemente dall'orientamento del parlante. Il sistema di riferimento **intrinseco** descrive la posizione in relazione all'orientamento dell'oggetto di riferimento: «la palla è di fronte al tavolo» fa riferimento al davanti del tavolo, non al davanti del parlante.

La ricerca di Levinson, Pederson e colleghi ha dimostrato che i parlanti di lingue con sistemi di riferimento diversi sviluppano strategie di memorizzazione spaziale diversamente orientate: i parlanti di lingue allocentriche mantengono più facilmente orientamenti assoluti anche in situazioni di rotazione corporea, mentre i parlanti di lingue egocentrine mostrano disorganizzazione della rappresentazione spaziale in seguito a rotazione. Questo suggerisce che il sistema di riferimento grammaticalizzato nella lingua co-evolve con o induce strategie cognitive di codificazione della memoria spaziale.

La relatività linguistica non opera tuttavia in modo uniforme su tutti i sistemi cognitivi. Il dibattito tra Paul Regier, Paul Kay e colleghi, da un lato, e la tradizione dei Whorfiani moderati dall'altro, ha chiarito che gli effetti categoriali del colore sono asimmetricamente distribuiti tra i due emisferi cerebrali: l'effetto di accelerazione nella discriminazione inter-categoriale è più pronunciato nell'emisfero sinistro — dominante per il linguaggio — e si attenua nell'emisfero destro, il che è consistente con l'ipotesi che il linguaggio interferisca online con la categorizzazione percettiva attraverso le risorse del sistema linguistico sinistro, ma non alteri la rappresentazione percettiva di basso livello gestita bilateralmente. La relatività linguistica, in questa lettura, non è un effetto sulla percezione grezza ma un effetto sul modo in cui le risorse linguistiche partecipano ai compiti di discriminazione online.

Formalizzando: sia L la struttura categoriale della lingua (la partizione dello spazio dei colori indotta dai termini cromatici), P la struttura percettiva di basso livello (la geometria dello spazio CIE), e C la struttura della discriminazione cognitiva nel compito. L'effetto Whorfiano moderato afferma che C non è determinato soltanto da P — come il Nativismo percettivo sostiene — ma è una funzione di P e L:

$C = f(P, L, \text{task_context})$

dove il contributo di L è modulato dal contesto del compito (linguistico vs. non-linguistico) e dalla disponibilità di risorse linguistiche (disturbate vs. libere dall'interferenza verbale). L'effetto categoriale non è una proprietà della percezione bruta ma della percezione mediata dalle risorse del sistema linguistico in condizioni di compito che richiedono categorizzazione.

III.5 — Il problema dell'acquisizione del linguaggio: il teorema di Gold, la povertà dello stimolo e la statistica nel corpus

Nessun aspetto delle scienze cognitive ha prodotto una controversia più duratura e più carica di implicazioni filosófiche del problema dell'acquisizione del linguaggio, che si pone all'incrocio tra la linguistica teorica, la psicologia dello sviluppo, la logica dell'apprendimento e la filosofia della mente. La questione fondamentale è quella che Chomsky ha denominato il Problema di Platone: come fa il bambino, a partire da dati linguistici finiti, rumorosi e strutturalmente non rappresentativi dell'intera grammatica della lingua, a convergere sulla grammatica corretta in pochi anni e in modo sostanzialmente uniforme tra bambini esposti alla stessa lingua? La risposta innatista afferma che il bambino non inferisce la grammatica dai dati ma la «attiva» sulla base di una dotazione linguistica innata, la **Grammatica Universale** (GU), che limita lo spazio delle grammatiche possibili a un insieme finito di strutture parametrizzate.

Il fondamento formale dell'argomento innatista è il celebre **teorema di Gold** (1967), che stabilisce un risultato di impossibilità sull'apprendimento di linguaggi da esempi positivi. Un **apprendente** è una funzione f che, dato un testo T_t — una sequenza finita di stringhe campionate dalla lingua target — produce un'ipotesi di grammatica $f(T_t)$. Un apprendente identifica nel limite un linguaggio L da un testo T se esiste un indice t_0 tale che per ogni $t > t_0$, $f(T_t)$ è una grammatica per L — cioè f converge eventualmente e definitivamente alla grammatica corretta. Una classe di linguaggi $\mathcal{L}$ è **identificabile nel limite** se esiste un apprendente che identifica nel limite ogni linguaggio in $\mathcal{L}$ da ogni testo di quel linguaggio.

Il risultato fondamentale di Gold è il seguente: qualsiasi classe di linguaggi che contenga tutti i linguaggi finiti (tutti gli insiemi finiti di stringhe) e almeno un linguaggio infinito non è identificabile nel limite da esempi positivi soltanto. Formalmente: sia $\mathcal{L}$ una classe tale che per ogni insieme finito S di stringhe, il linguaggio $L_S = S \in \mathcal{L}$, e sia $L_\infty \in \mathcal{L}$ un linguaggio infinito. Allora non esiste un apprendente f che identifichi nel limite tutti i linguaggi in $\mathcal{L}$ da testi. La dimostrazione è per costruzione avversariale: si costruisce un testo per L_∞ che, ad ogni stadio, inganna f nel ritenere che il linguaggio target sia un linguaggio finito, e al momento della convergenza si aggiunge una nuova stringa che obbliga f a aggiornare la propria ipotesi.

Poiché la classe dei linguaggi naturali include sia linguaggi finiti (ogni lingua naturale ha infinite frasi, ma ogni sistema che un bambino potrebbe ipoteticamente acquisire include strutture finito-infinite) sia linguaggi infiniti, e

poiché i bambini apprendono il linguaggio da esempi positivi senza feedback sistematico sugli errori (i genitori non correggono sistematicamente le costruzioni agrammaticali), il teorema di Gold implica che l'acquisizione del linguaggio da esempi positivi soli è, in linea di principio, impossibile per qualsiasi classe abbastanza ricca. La conclusione chomskyana è che la mente del bambino non impara nel senso ordinario ma è già equipaggiata con una conoscenza strutturale della grammatica possibile.

Il **Programma Principi e Parametri**, sviluppato da Chomsky e dai suoi collaboratori negli anni Ottanta, formalizza la GU come un sistema di principi universali (validi per tutte le lingue) e parametri binari (valori che variano tra le lingue e che il bambino imposta sulla base dell'input). Un principio universale è la condizione di località del movimento sintattico, che vieta certi spostamenti a lunga distanza; un parametro è la posizione della testa (Head Parameter): le lingue differiscono per il fatto che la testa di un sintagma preceda o segua il proprio complemento, e questa differenza spiega la distinzione tra lingue SOV (come il giapponese e il turco) e lingue SVO (come l'italiano e l'inglese). Con un numero limitato di parametri, si può generare la diversità osservata nelle strutture grammaticali delle lingue del mondo.

Il **Programma Minimalista** (Chomsky, 1995 e seguenti) riduce la GU al minimo: la sola operazione di base è Merge, che combina due oggetti sintattici α e β nell'insieme $\{\alpha, \beta\}$ (External Merge su oggetti distinti) o ri-combina un elemento già presente nella derivazione in una posizione sintattica diversa (Internal Merge, equivalente al «movimento» delle grammatiche precedenti). Tutta la struttura sintattica è generata per applicazione ricorsiva di Merge, e i parametri residui del programma precedente vengono ricondotti a proprietà del lessico piuttosto che della sintassi. La complessità sintattica emerge dalla semplicità computazionale di Merge.

L'argomento della **povertà dello stimolo** (APS) è la versione empirica del teorema di Gold: si identifica una struttura grammaticale specifica — un principio della GU — e si mostra che i bambini la acquisiscono correttamente sebbene i dati a cui sono esposti non contengano informazione sufficiente per inferirla induttivamente. Il caso classico è la formazione delle interrogative con movimento di testa: in italiano e in inglese, la frase interrogativa sì/no si forma spostando l'ausiliare in posizione iniziale. Per la frase «Il ragazzo che è simpatico è arrivato», l'interrogativa è «È arrivato il ragazzo che è simpatico?» e non «*È il ragazzo che arrivato è simpatico?». La regola della struttura superficiale (sposta il primo ausiliare) darebbe il risultato sbagliato; la regola dipendente dalla struttura (sposta l'ausiliare della proposizione principale) dà il risultato corretto. I bambini non producono mai la forma sbagliata, sebbene questo tipo di frase — con una relativa che interrompe il soggetto — sia estremamente raro nell'input. La conclusione dell'argomentazione chomskyana è che i bambini non stanno imparando la regola dall'input ma applicano un principio della GU che vieta le regole dipendenti dalla linearità.

La sfida più seria all'innatismo linguistico è venuta dalla ricerca sull'**apprendimento statistico**, inaugurata da Jenny Saffran, Richard Aslin e Elissa Newport (1996). I neonati di 8 mesi, esposti per soli 2 minuti a un flusso

continuo di sillabe senza pause o altri segnali acustici di confine, mostrano una preferenza per sequenze con alta probabilità di transizione statistica (*transitional probability*) tra sillabe, rispetto a sequenze con bassa probabilità di transizione. La probabilità di transizione condizionale da una sillaba X a una sillaba Y è definita come:

$$TP(Y|X) = P(XY) / P(X) = n(XY) / n(X)$$

dove n(XY) è la frequenza del bigramma XY e n(X) è la frequenza della sillaba X. Nelle sequenze sperimentali, le probabilità di transizione all'interno delle «parole» (sequenze fisse) erano di 1.0, mentre le probabilità di transizione tra «parole» erano di 0.33. I neonati distinguevano statisticamente tra sequenze ad alta e bassa probabilità di transizione, dimostrando una capacità di apprendimento statistico straordinariamente rapida e implicita.

Questo risultato ha riaperto il dibattito sull'APS: se i bambini sono in grado di estrarre regolarità statistiche dal flusso linguistico con tale efficienza, forse le strutture grammaticali sono in parte inferibili dall'input statistico, e l'argomento della povertà dello stimolo è meno cogente di quanto Chomsky sostenesse. La risposta innatista è che l'apprendimento statistico delle probabilità di transizione spiega l'acquisizione della struttura fonologica superficiale — la segmentazione del flusso in unità lessicali — ma non le strutture sintattiche astratte, dipendenti dalla struttura gerarchica delle frasi, che non sono direttamente codificate in probabilità di transizione lineari.

Il dibattito è, in larga misura, irrisolto. Un approccio riconciliatorio — sviluppato da ricercatori come Morten Christiansen, Nick Chater e Simon Kirby — propone che le strutture grammaticali universali non siano innate nel senso di prespecificate geneticamente, ma siano il prodotto dell'interazione tra il sistema cognitivo umano (con le sue capacità di apprendimento statistico, di elaborazione sequenziale, di compressione dell'informazione) e la pressione evolutiva della comunicazione sulla forma delle lingue nel corso della trasmissione culturale. Le lingue si adattano alle capacità cognitive degli apprendenti, e le capacità cognitive si adattano alle lingue — una coevoluzione che produce la convergenza osservata senza richiedere né un innatismo forte né un apprendimento puramente statistico. La struttura della lingua è allora né interamente innata né interamente indotta dall'input, ma co-determinata dall'interazione tra la cognizione incarnata del parlante e la tradizione linguistica della comunità.

III.6 — Semantica cognitiva: costruzioni, spazi mentali e la mescolanza concettuale

Il programma della semantica cognitiva — che si dispiega nei lavori di Ronald Langacker, Charles Fillmore, Gilles Fauconnier e Mark Turner a partire dagli anni Ottanta — parte da una premessa radicalmente diversa da quella della semantica formale montagoniana: la grammatica non è un sistema autonomo di regole formali che opera su rappresentazioni astratte indipendenti dal significato, ma è una struttura di unità simboliche convenzionalizzate, cioè di coppie forma-significato, in cui il significato è concettuale nel senso della semantica

cognitiva e la forma è fonologica. Non esiste una sintassi autonoma: ogni struttura grammaticale è simultaneamente struttura formale e struttura di significato.

Langacker denomina questo programma **Grammatica Cognitiva** (*Cognitive Grammar*): le unità grammaticali sono **unità simboliche** — coppie ⟨F, S⟩ dove F è il polo fonologico (la struttura di forma) e S è il polo semantico (la struttura di significato). Il morfema «cane» è un'unità simbolica che accoppia la sequenza fonemica /kɑnɛ/ con il concetto CANE. La preposizione «su» accoppia la sequenza fonemica /su/ con uno schema topologico di traettore posizionato superiormente rispetto a un punto di riferimento. La differenza tra le categorie grammaticali — sostantivi, verbi, aggettivi, preposizioni — non è definita dal tipo di complementazione sintattica che selezionano ma dal tipo di struttura concettuale che profilano: i sostantivi profilano cose (regioni in qualche dominio concettuale), i verbi profilano relazioni temporalmente distribuite, gli aggettivi profilano proprietà non temporalmente distribuite, le preposizioni profilano relazioni spaziali.

Il concetto di **profilo** è fondamentale nella Grammatica Cognitiva: ogni espressione linguistica richiama alla mente una struttura concettuale più ampia — la **base** o il **dominio di base** — e designa (profila) una struttura specifica all'interno di quella base. Il termine «lunedì» profila un giorno della settimana, ma lo fa nell'ambito del dominio concettuale del ciclo settimanale; il termine «fratello» profila un individuo di sesso maschile nell'ambito del dominio della struttura familiare; il termine «acquirente» profila il ruolo di chi acquista nell'ambito del frame della transazione commerciale. Senza la base, il profilo è semanticamente incompleto: si può capire «lunedì» soltanto sullo sfondo del ciclo settimanale.

La **Frame Semantics** di Charles Fillmore radicalizza questa idea: il significato delle parole non è una lista di tratti semantici o una posizione in uno spazio concettuale, ma l'evocazione di un **frame** — una struttura di conoscenza conventzionalizzata che descrive un tipo di situazione o evento con i suoi partecipanti, ruoli e relazioni caratteristici. Il frame della TRANSAZIONE COMMERCIALE è strutturato da ruoli: VENDITORE, ACQUIRENTE, MERCE, DENARO, TRASFERIMENTO. I termini «vendere», «comprare», «pagare», «costare», «spendere» evocano tutti lo stesso frame transazionale, ma profilano prospettive diverse su di esso: «vendere» profila l'azione dal punto di vista del venditore, «comprare» dal punto di vista dell'acquirente, «costare» dalla prospettiva del prezzo della merce. La comprensione di qualsiasi termine richiede l'attivazione dell'intero frame di sfondo, non soltanto il profilo designato.

Il programma FrameNet, sviluppato da Fillmore, Baker e colleghi presso ICSI Berkeley, ha sistematizzato questo approccio in un archivio lessicale computazionale in cui ogni unità lessicale è annotata con il frame che evoca e con la mappatura tra i suoi complementi sintattici e i ruoli del frame (Frame Elements). Il verbo «accusare» evoca il frame ACCUSATION con ruoli ACCUSATORE, ACCUSATO e COMPORTAMENTO; i complementi sintattici di «accusare» in una frase concreta si mappano su questi ruoli secondo convenzioni costruzionali regolari.

La **Grammatica Costruzionale** di Adele Goldberg (*Constructions: A Construction Grammar Approach to Argument Structure*, 1995) ha esteso il programma: le costruzioni grammaticali — non solo le unità lessicali ma anche i pattern sintattici astratti come la costruzione ditransitiva (Soggetto Verbo Oggetto Oggetto, es. «Maria diede un libro a Giovanni») e la costruzione con causativa risultativa (Soggetto Verbo Oggetto Predicato risultativo, es. «Il martello rese piatta la superficie») — sono esse stesse unità simboliche convenzionalizzate con significato costruzionale proprio. Il significato di «Maria diede a Giovanni la febbre» non è derivabile dal significato di «dare» — che normalmente richiede un'azione fisica di trasferimento — ma dalla costruzione ditransitiva, che contribuisce il suo significato autonomo di TRASFERIMENTO CAUSATO. La costruzione ha significato indipendente dal verbo che la occupa, e questo significato interagisce con il significato del verbo per produrre il significato della frase complessiva.

La teoria degli **Spazi Mentali** di Gilles Fauconnier (*Mental Spaces*, 1985) fornisce un quadro dinamico per la semantica del discorso che supera le limitazioni della semantica proposizionale intensionale montagoniana. Gli spazi mentali sono strutture parziali costruite dinamicamente durante la comprensione del discorso, che servono come domini separati di riferimento e predicazione. Ogni spazio mentale è una struttura $\langle E, R, \Phi \rangle$ dove E è un insieme di elementi (individui nello spazio), R è un insieme di relazioni tra elementi, e Φ è la funzione di connettore che collega elementi di spazi diversi: per ogni elemento b nello spazio figlio (target space), $\Phi(b) = a$ è l'elemento controparte di b nello spazio genitore (parent space). Il discorso procede costruendo spazi, aprendo nuovi spazi tramite costruttori di spazio (operatori linguistici come «nel film», «in the dream», «nel 1945», «se», «supponiamo che»), e gestendo le connessioni tra i controparte negli spazi.

La costruzione di spazi mentali spiega fenomeni referenziali altrimenti difficili da trattare. L'enunciato «Nel film, Charlot sposa la ballerina» apre uno spazio cinematografico S_film distinto dallo spazio della realtà attuale S_R. L'elemento Charlot in S_film è connesso da Φ all'attore Charlie Chaplin in S_R; l'elemento «la ballerina» in S_film non ha necessariamente un controparte in S_R. I pronomi e i sintagmi nominali possono essere referenzialmente ancorati a spazi diversi, e il discorso può navigare tra spazi attraverso i connettori, creando strutture complesse di riferimento incrociato.

La **Teoria dell'Integrazione Concettuale** (Conceptual Blending) di Fauconnier e Turner (*The Way We Think*, 2002) estende gli spazi mentali introducendo il meccanismo della mescolanza (*blending*): oltre agli spazi di input (tipicamente due) e allo spazio generico (che cattura la struttura comune agli input), esiste uno **spazio blended** che eredita struttura selettivamente dagli input e sviluppa **struttura emergente** — struttura non presente in nessuno degli input — attraverso l'interazione delle strutture ereditate. La struttura emergente è ciò che rende il blend intellettualmente produttivo: non è una mera somma degli input ma una configurazione nuova con proprietà proprie.

La struttura categoriale del blend può essere formalizzata nei termini di un **pushout** nella categoria degli spazi mentali. Dati due spazi di input I_1 e I_2 con

mappature $g_1\colon G \to I_1$ e $g_2\colon G \to I_2$ dallo spazio generico G, il blend B è il colimite di questo diagramma — l'oggetto universale B con morfismi $f_1\colon I_1 \to B$ e $f_2\colon I_2 \to B$ tali che $f_1 \circ g_1 = f_2 \circ g_2$. Formalmente:

$$G \xrightarrow{g_1} I_1 \quad | \quad | \; g_2 \quad f_1 \; | \quad \downarrow \quad \downarrow \; I_2 \xrightarrow{f_2} B$$

Il blend B è il pushout di g_1 e g_2: contiene le proiezioni di entrambi gli input identificate sullo spazio generico. La **struttura emergente** corrisponde alle proprietà di B che non sono l'immagine di alcun elemento di $I_1 \sqcup I_2$ attraverso f_1 o f_2, ma emergono dall'interazione delle strutture ereditate nel blend — dalla «pressione» delle strutture concorrenti che si integrano in una configurazione coerente. L'analisi categoriale cattura perché il blend non è riducibile ai suoi input: il colimite ha proprietà che nessuno dei componenti del diagramma possiede.

III.7 — Intenzionalità, linguaggio del pensiero e fondamenti teleosemantici del contenuto mentale

Franz Brentano, nelle *Lezioni di psicologia empirica* del 1874, propose quella che è diventata probabilmente la caratterizzazione più influente del mentale nella filosofia contemporanea: ogni fenomeno psichico è caratterizzato dall'**intenzionalità** — dalla «direzione verso un oggetto» (*Gerichtetsein auf ein Objekt*) o dall'«inerenza immanente di un contenuto». Percepire è percepire qualcosa; credere è credere qualcosa; volere è volere qualcosa; temere è temere qualcosa. Questa «aboutness» — il fatto che ogni atto mentale si orienta verso un oggetto che può anche non esistere (posso pensare a un unicorno anche in assenza di unicorni) — è, secondo Brentano, il segno distintivo del mentale rispetto al fisico: i fenomeni fisici non hanno questa direzione intenzionale intrinseca.

Edmund Husserl, partendo dall'ontologia intenzionale di Brentano ma orientandola verso una fenomenologia trascendentale, introduce la distinzione fondamentale tra la **noesi** (l'atto di intendere, la struttura dell'atto mentale in quanto atto) e il **noema** (il contenuto inteso, l'oggetto «come inteso» — non l'oggetto reale ma il modo in cui si presenta alla coscienza). Il noema è strutturato: ha un «nucleo noemático» che determina il referente attraverso un sistema di predicati intenzionali, e ha un «carattere tetico» che ne determina la modalità di presenza (percepito, ricordato, immaginato, giudicato vero). La fenomenologia husserliana è così una teoria del contenuto intenzionale: il significato di un atto mentale è il suo noema, la struttura che determina a quale oggetto (se esiste) l'atto si riferisce e come lo presenta.

La domanda filosofica che si pone è come l'intenzionalità si origina nei sistemi fisici. Per Husserl, essa appartiene alla struttura trascendentale della coscienza; per il fisicalismo, deve essere spiegata in termini fisici o funzionali. Jerry Fodor ha proposto l'ipotesi più influente: la **Language of Thought Hypothesis** (LOTH), esposta in *The Language of Thought* (1975) e difesa per decenni successivi. L'ipotesi sostiene che gli stati mentali proposizionali — credenze, desideri, intenzioni — sono relazioni tra un soggetto e formule di un linguaggio interno di tipo,

denominato Mentalese. Mentalese ha una sintassi — le formule sono oggetti strutturalmente complessi — e una semantica — le formule hanno contenuto, si riferiscono a oggetti nel mondo e hanno condizioni di verità. I processi cognitivi sono processi computazionali — cioè processi che operano sulle formule di Mentalese in virtù della loro struttura sintattica, producendo transizioni tra stati mentali in modi che rispettano le relazioni semantiche tra i contenuti.

Il nucleo dell'argomentazione di Fodor in favore della LOTH è l'argomento dalla **sistematicità** e dall'**operatività produttiva** del pensiero. La sistematicità del pensiero afferma che le capacità cognitive non sono casualmente indipendenti: se un agente è in grado di pensare «Giovanni ama Maria», è necessariamente in grado di pensare «Maria ama Giovanni». Questa non è una correlazione contingente ma una necessità strutturale: i due pensieri condividono i componenti (GIOVANNI, MARIA, AMARE) e differiscono soltanto per le relazioni tra i componenti, il che implica che i componenti sono individualmente rappresentati e combinabili in modi sistematici. La struttura composizionale di Mentalese spiega questa sistematicità: la formula AMARE(giovanni, maria) e la formula AMARE(maria, giovanni) condividono i componenti sintattico-semantici ma li dispongono in modo diverso, e un sistema che può costruire la prima può costruire anche la seconda.

La **produttività** del pensiero — il fatto che un agente competente può pensare un numero potenzialmente illimitato di pensieri distinti — richiede allo stesso modo che il pensiero abbia struttura compositiva ricorsiva: non si possono memorizzare tutti i pensieri possibili, si devono invece combinare un numero finito di componenti elementari attraverso un sistema ricorsivo di regole. Un linguaggio del pensiero con regole ricorsive genera un insieme infinito di formule, esattamente come la sintassi ricorsiva del linguaggio naturale genera un insieme infinito di frasi da un lessico finito.

La critica più sistematica alla LOTH è venuta dal connessionismo, che propone di spiegare i medesimi fenomeni cognitivi senza postulare rappresentazioni simboliche strutturate. La risposta di Fodor e Pylyshyn (1988) — in «Connectionism and Cognitive Architecture: A Critical Analysis» — è che i modelli connessionistici possono implementare sistemi sistematici e produttivi soltanto se, al livello dell'implementazione, realizzano qualcosa che ha la struttura di un linguaggio del pensiero. Un sistema che è genuinamente sistematico deve avere componenti strutturali distinti per GIOVANNI, MARIA e AMARE, e un meccanismo per combinarli in modi diversi: questo è precisamente ciò che la LOTH postula. Se il sistema connessionista non ha questa struttura implicita, non sarà genuinamente sistematico ma soltanto approssimativamente.

Fred Dretske ha proposto, in *Knowledge and the Flow of Information* (1981) e in *Explaining Behavior* (1988), un approccio naturalistico al contenuto mentale basato sulla nozione informazionale. L'idea di base è che il contenuto di una rappresentazione interna è determinato dall'informazione che essa porta nel senso informazionale-teorico: una struttura r porta l'informazione che P (in un canale C) se e soltanto se la probabilità condizionale che P sia il caso dato r e il canale C è uguale a 1:

$P(P \mid r, C) = 1$

Il significato naturale (*natural meaning*) di un segnale — nel senso di Grice — è l'informazione che esso porta in condizioni normali di trasmissione: il fumo significa naturalmente fuoco perché porta l'informazione che c'è fuoco. Dretske propone di costruire la semantica intenzionale — il significato non naturale degli stati mentali — come caso speciale del significato naturale, dove le «condizioni normali» non sono le condizioni fisiche ordinarie ma le condizioni di design funzionale del sistema cognitivo.

La **Teleosemantica** di Ruth Millikan (*Language, Thought, and Other Biological Categories*, 1984) radicalizza l'approccio dretskiano: il contenuto di una rappresentazione mentale è determinato dalla funzione biologica propria (*proper function*) della struttura che la produce — la funzione per cui quella struttura è stata selezionata nell'evoluzione. Un sistema percettivo che produce uno stato r è stato selezionato perché produceva r in presenza di P e perché questo correlazione permetteva all'organismo di comportarsi in modi adattativi. Il contenuto di r è P — l'oggetto nel mondo la cui presenza causava la produzione di r nelle condizioni in cui il sistema è stato selezionato. Il contenuto non è determinato da ciò che r rappresenta de facto (che potrebbe essere qualsiasi cosa correlata causalmente a r), ma da ciò che r ha la funzione biologica di rappresentare — dalle condizioni in cui il sistema rappresentante ha ottenuto vantaggi adattativi grazie al produrre r.

Il contenuto teleosemantico è, in questo senso, normativo: la rappresentazione ha il contenuto P anche quando le condizioni di produzione si discostano dalle condizioni normali di selezione e la rappresentazione è quindi errata. Questo risolve il problema della **mala rappresentazione** (*misrepresentation*): un sistema informazionale puro che si basa sulla covariazione causale non può spiegare l'errore (se r co-varia causalmente con X, allora r «significa» X e non può mai essere errato rispetto a X); la teleosemantica spiega l'errore come il caso in cui r è prodotta fuori dalle condizioni normali, ovvero quando r è prodotta pur in assenza di P — le condizioni per cui il sistema è stato selezionato.

III.8 — Connessionismo e cognizione: reti neurali come modelli dell'elaborazione distribuita

Il connessionismo — denominazione che racchiude un insieme di approcci alla modellazione della cognizione che fa riferimento a reti di unità semplici interconnesse, ispirate ai neuroni biologici — è il programma che ha più radicalmente sfidato il paradigma simbolico dell'intelligenza artificiale classica e, con esso, la LOTH e il cognitivismo computazionale. La sua storia è scandita da cicli di entusiasmo e disincanto che riflettono tanto le difficoltà tecniche quanto i malintesi filosófici sul rapporto tra i modelli connessionisti e la cognizione.

La formalizzazione pionieristica di Warren McCulloch e Walter Pitts (1943) introduce il **neurone formale**: un'unità logica che prende in ingresso segnali binari $x_1, ..., x$, li pondera con pesi $w_1, ..., w$, e produce un output binario $y = 1$ se la somma pesata supera una soglia θ, $y = 0$ altrimenti:

$y = H(\Sigma_i w_i x_i - \theta)$

dove H è la funzione di Heaviside. McCulloch e Pitts dimostrano che reti di neuroni formali, con connessioni opportune, possono computare qualsiasi funzione booleana — sono Turing-complete. Il **Percettrone** di Frank Rosenblatt (1958) aggiunge l'apprendimento: i pesi sono aggiornati iterativamente sulla base dell'errore tra l'output prodotto e l'output desiderato. Il teorema di convergenza del percettrone afferma che se le classi sono **linearmente separabili** — se esiste un iperpiano nello spazio degli input che le separa — allora l'algoritmo di aggiornamento converge in un numero finito di passi a una soluzione.

La limitazione identificata da Marvin Minsky e Seymour Papert in *Perceptrons* (1969) è tecnica ma filosofica mente significativa: i percettroni a singolo strato non possono computare la funzione XOR, perché XOR non è linearmente separabile nello spazio degli input booleani. La pubblicazione di questo risultato contribuì al primo «inverno dell'IA» connessionista. La ripresa avvenne con la riscoperta e la diffusione dell'algoritmo di **retropropagazione dell'errore** (*backpropagation*) nelle reti multi-strato, formalizzato indipendentemente da vari autori e reso popolare da David Rumelhart, Geoffrey Hinton e Ronald Williams (1986).

In una rete *feedforward* multi-strato con uno strato di input, L strati nascosti e uno strato di output, le attivazioni si propagano in avanti: per ogni strato l (con l = 1, ..., L+1), l'attivazione del neurone j è:

$a^l_j = \sigma(\Sigma_i w^l_{ji} \cdot a^{l-1}_i + b^l_j)$

dove σ è la funzione di attivazione (sigmoide $\sigma(x) = 1/(1+e^{-x})$, o tanh, o ReLU = $\max(0,x)$), w^l_{ji} è il peso della connessione dal neurone i dello strato l-1 al neurone j dello strato l, e b^l_j è il bias. L'errore dell'output è calcolato attraverso una funzione di costo E (tipicamente l'errore quadratico medio o la cross-entropia). La retropropagazione calcola i gradienti dell'errore rispetto ai pesi per ogni strato, applicando la regola della catena del calcolo differenziale all'indietro dalla funzione di costo agli strati iniziali:

$\partial E/\partial w^l_{ji} = \partial E/\partial a^l_j \cdot \partial a^l_j/\partial z^l_j \cdot \partial z^l_j/\partial w^l_{ji} = \delta^l_j \cdot a^{l-1}_i$

dove $\delta^l_j = \partial E/\partial z^l_j$ è il «segnale di errore» del neurone j nello strato l (denominato «delta»), e $z^l_j = \Sigma_i w^l_{ji} \cdot a^{l-1}_i + b^l_j$ è il preattivazione. Il delta si propaga all'indietro attraverso la rete:

$\delta^l_j = \sigma'(z^l_j) \cdot \Sigma_k w^{l+1}_{kj} \cdot \delta^{l+1}_k$

dove σ' è la derivata della funzione di attivazione. I pesi sono aggiornati nella direzione del gradiente negativo:

$w^l_{ji} \leftarrow w^l_{ji} - \eta \cdot (\partial E/\partial w^l_{ji})$

con η il tasso di apprendimento. Questa procedura di discesa del gradiente stocastica converge — sotto opportune condizioni sulla funzione di costo e sulle dimensioni della rete — verso un minimo locale dell'errore.

Il significato filosofico del connessionismo per la teoria della cognizione è dibattuto. Paul Smolensky ha proposto, in «On the Proper Treatment of Connectionism» (1988), che i modelli connessionistici non debbano essere interpretati al livello della simulazione computazionale del livello simbolico — come se implementassero un linguaggio del pensiero nascosto — ma come un livello di descrizione proprio, il **livello sub-simbolico**, in cui le unità di elaborazione non corrispondono a concetti o simboli ma a microcaratteristiche sub-concettuali. I concetti emergono come pattern di attivazione distribuiti sull'intera rete, non come entità localizzate in unità specifiche.

Smolensky ha poi proposto le **rappresentazioni tensoriali** (*Tensor Product Representations*, TPR) come strumento per formalizzare come i modelli connessionisti possano realizzare strutture composizionali senza postulare simboli locali. Una TPR di un'espressione strutturata è la somma dei prodotti tensoriali tra vettori di riempitivo (*filler vectors*) — che rappresentano il contenuto — e vettori di ruolo (*role vectors*) — che rappresentano la posizione strutturale:

$$T = \Sigma_i \, f_i \otimes r_i$$

dove f_i è il vettore del riempitivo dell'i-esimo costituente e r_i è il vettore del ruolo corrispondente, e $\otimes$ è il prodotto tensoriale. Per la struttura «Giovanni ama Maria» con ruoli r_soggetto, r_verbo, r_oggetto:

$$T = v_Giovanni \otimes r_soggetto + v_ama \otimes r_verbo + v_Maria \otimes r_oggetto$$

Il contenuto del riempitivo in un ruolo specifico può essere estratto moltiplicando per lo pseudo-inverso del vettore del ruolo: $\tilde{f}_i = T \cdot r_i^{+} \approx f_i$ (con piccola interferenza dagli altri componenti se i vettori di ruolo sono quasi-ortogonali). Questo dimostra che le TPR forniscono rappresentazioni distribuite e composizionali: la sistematicità e la produttività che Fodor e Pylyshyn richiedevano emergono dalla struttura tensoriale senza richiedere entità simboliche localizzate.

Il dibattito tra cognitivismo simbolico e connessionismo non si è risolto ma si è trasformato: con l'avvento del deep learning — reti profonde con molti strati nascosti, addestrate su grandi quantità di dati — e con la dimostrazione che queste reti raggiungono prestazioni straordinarie in compiti di percezione, comprensione del linguaggio e generazione, la domanda non è più se le reti neurali possono competere con i sistemi simbolici ma piuttosto quale tipo di rappresentazione e di processo cognitivo implementino le reti profonde, e se questa implementazione abbia qualcosa di rilevante da dire sulla cognizione umana.

Il **deep learning** — termine che designa le reti neurali con molti strati nascosti addestrate su grandi quantità di dati con tecniche di ottimizzazione stocastica — ha prodotto, nel secondo decennio del ventunesimo secolo, risultati che hanno trasformato non solo la ricerca applicata ma le concezioni teoriche della cognizione. La dimostrazione che reti con architetture relativamente semplici — strati convoluzionali, strati fully-connected, meccanismi di attenzione — addestrate su milioni di esempi raggiungono prestazioni superiori agli esseri umani in compiti di riconoscimento di immagini, di traduzione automatica, di

giochi da tavolo, ha posto una sfida alla concezione della specializzazione cognitiva: se una sola architettura generale impara così tanto da così poco specifica progettazione, forse la specificità cognitiva emerge dall'apprendimento piuttosto che dall'architettura innata.

La questione filosofica che il deep learning pone con urgenza rinnovata è quella delle **rappresentazioni interne**: cosa apprendono le reti profonde, e quelle rappresentazioni hanno qualcosa che assomiglia alle strutture concettuali che la psicologia cognitiva identifica? La tecnica della **feature visualization** — che massimizza l'attivazione di unità individuali o insiemi di unità attraverso la sintesi di input ottimali — ha rivelato che le reti convoluzionali addestrate su immagini sviluppano, negli strati intermedi, rappresentazioni che assomigliano a caratteristiche percettive di interesse crescente: negli strati bassi, bordi e gradienti; negli strati intermedi, texture e forme; negli strati superiori, configurazioni di oggetti e parti di oggetti. Questa progressione di complessità crescente è sorprendentemente analoga alla gerarchia del sistema visivo ventrale — V1, V2, V4, IT — descritta dalla neuroscienze computazionali, e ha alimentato l'ipotesi che le reti profonde realizzino algoritmi analoghi a quelli implementati dalla corteccia visiva, sebbene attraverso substrati e meccanismi diversi.

III.8bis — L'elaborazione predittiva e il principio di energia libera: la cognizione come inferenza bayesiana incarnata

L'esame del connessionismo condotto nella sezione III.8 si è chiuso sull'interrogazione se le rappresentazioni interne delle reti neurali profonde abbiano qualcosa che assomigli alle strutture concettuali identificate dalla psicologia cognitiva, aprendo la domanda di quale tipo di elaborazione le reti implementino e se questa elaborazione abbia implicazioni per la comprensione della cognizione umana. Un'intera prospettiva teorica sulla natura della cognizione — maturata in modo sistematico a partire dagli anni Duemila attraverso il confluire di teorie neuroscientifiche, modelli bayesiani e filosofia della mente, eppure con radici risalenti alla seconda metà dell'Ottocento — è rimasta finora assente dall'analisi: il **paradigma dell'elaborazione predittiva** (*predictive processing*), nelle sue diverse formulazioni, costituisce forse il tentativo più ambizioso e più formalmente preciso di fornire un resoconto unificato della cognizione come processo di inferenza bayesiana attiva, e ha implicazioni dirette per le domande sul significato, sul radicamento e sull'intenzionalità che sono al centro di questa monografia.

La genealogia intellettuale del paradigma ha un'origine sorprendentemente lontana. Hermann von Helmholtz, nella sua monumentale opera sulla fisiologia ottica degli anni 1860, aveva proposto che la percezione fosse essenzialmente **inferenza inconscia** (*unbewusste Schlüsse*): il sistema percettivo non recepisce passivamente le informazioni sensoriali ma le interpreta attivamente, costruendo la migliore spiegazione causale possibile degli stimoli in arrivo sulla base delle conoscenze pregresse. La retina, per Helmholtz, non trasmette immagini al cervello; trasmette segnali — variazioni di intensità luminosa localizzate — che il sistema visivo trasforma in percezioni di oggetti tridimensionali, superfici, profondità e movimenti attraverso un processo inferenziale che incorpora

conoscenze apprese sull'ambiente, sulla fisica della luce e sulle strutture degli oggetti. La percezione dell'oggetto non è nel segnale sensoriale: è nel prodotto dell'inferenza che il sistema cognitivo compie a partire dal segnale, utilizzando un modello implicito del mondo.

Questa intuizione rimase a lungo in stato di latenza, assorbita dalla tradizione comportamentista e poi dall'approccio computazionale classico che concepiva la cognizione come elaborazione bottom-up di rappresentazioni. La sua ripresa sistematica in forme matematicamente precise avviene con il lavoro di Rajesh Rao e Dana Ballard pubblicato nel 1999 su *Nature Neuroscience*, che propone un modello di **codifica predittiva gerarchica** (*hierarchical predictive coding*) per il sistema visivo. Il modello di Rao e Ballard descrive la corteccia visiva come una gerarchia di livelli, ognuno dei quali mantiene un modello delle cause probabilistiche degli input che riceve dal livello inferiore. Ogni livello invia **predizioni top-down** al livello immediatamente inferiore, descrivendo ciò che si aspetta di ricevere dati i propri stati correnti. Il livello inferiore calcola la **discrepanza** (*prediction error*) tra le predizioni ricevute e gli stimoli effettivi, e trasmette soltanto questa discrepanza al livello superiore — non il segnale grezzo, ma la deviazione dal previsto. Il livello superiore aggiorna i propri stati interni per ridurre l'errore di predizione ricevuto, e invia predizioni aggiornate al livello inferiore. Il processo converge verso una rappresentazione dell'input che minimizza gli errori di predizione a tutti i livelli della gerarchia.

Questo schema è straordinariamente parsimonioso: invece di trasmettere informazioni ridondanti attraverso tutta la gerarchia, il sistema trasmette soltanto le deviazioni dal previsto — soltanto l'informazione che costituisce sorpresa rispetto al modello corrente. La codifica predittiva è dunque non soltanto un modello neuroscientificamente plausibile ma un principio di efficienza informativa: un sistema che predice accuratamente il proprio input deve trasmettere soltanto gli errori, e la dimensione degli errori è proporzionale alla sorpresa — all'improbabilità dell'input dato il modello. La connessione con il formalismo dell'informazione di Shannon è diretta: la sorpresa di un evento x è $-\log P(x)$, e minimizzare la sorpresa equivale a massimizzare la probabilità dell'input dati i modelli interni, ovvero a costruire modelli sempre più accurati del mondo.

Karl Friston ha generalizzato questo schema locale, sviluppato da Rao e Ballard per il solo sistema visivo, in un principio unificato della cognizione denominato **principio di energia libera** (*free energy principle*). Il principio, elaborato in una serie di articoli tra il 2005 e il 2010, afferma che ogni sistema biologico che resiste alla propria dissipazione — che mantiene la propria organizzazione nel tempo nonostante il secondo principio della termodinamica — deve agire in modo da minimizzare una quantità denominata **energia libera variazionale**. L'energia libera variazionale F è una funzione dei segnali sensoriali o (le osservazioni) e degli stati interni μ del sistema che approssima la sorpresa del segnale sensoriale:

$$F(o, \mu) = -\langle \log p(o, x) \rangle_q + \langle \log q(x; \mu) \rangle_q$$

dove x sono gli stati nascosti del mondo (le cause delle osservazioni), $q(x; \mu)$ è la distribuzione approssimata sulle cause — il modello interno del sistema

parametrizzato dagli stati interni μ — e $p(o, x)$ è il modello generativo del sistema — la distribuzione congiunta sulle osservazioni e le cause che il sistema possiede implicitamente. Questa espressione si scompone in due componenti di significato distinto:

$$F = KL[q(x; \mu) \parallel p(x|o)] - \log p(o)$$

dove $KL[q \parallel p]$ è la divergenza di Kullback-Leibler tra la distribuzione approssimata q e la vera distribuzione posteriore $p(x|o)$, e $-\log p(o)$ è la sorpresa del segnale sensoriale. Poiché la divergenza KL è sempre non negativa, l'energia libera è sempre un limite superiore della sorpresa: $F \geq -\log p(o)$. Il sistema può minimizzare F in due modi fondamentalmente diversi. Il primo consiste nel modificare gli stati interni μ in modo da avvicinare $q(x; \mu)$ alla vera distribuzione posteriore $p(x|o)$ — riducendo il termine KL e affinando la rappresentazione interna delle cause del segnale. Questa è la **percezione**, nel senso di inferenza sulle cause del segnale sensoriale. Il secondo consiste nel modificare i segnali sensoriali o attraverso l'azione — agendo sul mondo in modo da rendere le osservazioni più coerenti con le predizioni del modello interno, riducendo così la sorpresa $-\log p(o)$. Questo è l'**azione**, nel senso di comportamento che riduce la discrepanza tra le predizioni e le osservazioni effettive.

Il principio di energia libera unifica dunque percezione e azione sotto un'unica formulazione: entrambe sono strategie per minimizzare la discrepanza tra il modello interno del sistema e gli stati del mondo. La percezione minimizza l'energia libera aggiornando il modello; l'azione la minimizza cambiando il mondo. Friston denomina questo secondo modo di minimizzare l'energia libera **inferenza attiva** (*active inference*): il sistema non soltanto inferisce gli stati del mondo dai segnali sensoriali, ma modifica attivamente i segnali sensoriali agendo sul mondo. Questa formulazione dà conto in modo naturale del comportamento diretto agli obiettivi: le predizioni del sistema non riguardano soltanto ciò che si aspetta di percepire ma ciò che si aspetta di percepire se agisce in certi modi — le **predizioni propriocettive** che specificano gli stati corporei attesi durante l'azione. L'azione è essenzialmente la realizzazione di queste predizioni: il sistema muscolare implementa le azioni che portano gli stati del corpo ad allinearsi con le predizioni propriocettive, minimizzando l'errore di predizione propriocettivo.

La **ponderazione della precisione** (*precision weighting*) è il meccanismo attraverso cui l'elaborazione predittiva implementa l'**attenzione**. Non tutti gli errori di predizione hanno lo stesso peso nell'aggiornamento del modello: un segnale sensoriale proveniente da una sorgente affidabile — in condizioni di buona illuminazione, in assenza di rumore — dovrebbe ricevere più peso di un segnale incerto, perché porta più informazione sui veri stati del mondo. La precisione di un segnale è l'inverso della sua varianza: un segnale ad alta precisione è un segnale a bassa varianza, affidabile e informativo. Nell'architettura gerarchica della codifica predittiva, i messaggi bottom-up (gli errori di predizione) sono ponderati dalla loro precisione stimata, e i messaggi top-down (le predizioni) sono ponderati dalla fiducia nel modello generativo. L'attenzione, in questo quadro, è la modulazione della precisione dei messaggi di errore provenienti da specifiche sorgenti: orientare l'attenzione verso uno stimolo equivale ad aumentare la precisione attribuita agli errori di predizione generati da quel stimolo, dando loro

più peso nell'aggiornamento del modello. Questo meccanismo implementa formalmente il **filtraggio attentivo** — la selezione di quali errori di predizione devono essere trasmessi ai livelli superiori e quali possono essere soppressi — senza richiedere un sistema attenzionale separato dal sistema di codifica: l'attenzione è endogena alla struttura dell'elaborazione predittiva.

Le implicazioni del paradigma predittivo per la comprensione del linguaggio sono state esplorate con crescente sistematicità a partire dalla metà degli anni Duemila, trovando nel componente N400 dell'elettroencefalogramma (EEG) il correlato neurofisiologico più direttamente rilevante. L'N400 è una componente negativa del potenziale evocato che si registra in risposta a parole semanticamente anomale o improbabili nel contesto — la risposta alla parola «martello» nella frase «Ho bevuto una tazza di martello» è molto più ampia di quella alla parola «caffè» nella frase «Ho bevuto una tazza di caffè». La scoperta di Kutas e Hillyard nel 1980 che l'N400 scala con l'anomalia semantica ha orientato decenni di ricerca sulla comprensione del linguaggio. L'interpretazione predittiva dell'N400 è che esso riflette l'errore di predizione generato dalla parola effettiva rispetto alla parola predetta dal sistema di comprensione: una parola altamente predetta genera un errore piccolo e dunque un N400 ridotto; una parola imprevista genera un errore grande e un N400 ampio. Il sistema di comprensione del linguaggio non aspetta di ricevere la parola successiva per elaborarla: anticipa, sulla base del contesto morfosintattico e semantico-pragmatico, quale parola (o classe di parole) è probabile che arrivi, e si prepara a elaborarla. La comprensione è predizione; il significato di un enunciato si dispiega come una sequenza di aggiornamenti predittivi che incorporano ogni nuova parola nella struttura di anticipazione già costruita.

Questa formulazione — il significato come struttura predittiva — ha conseguenze filosófiche di prima grandezza per le domande che questa monografia persegue. Nella prospettiva dell'elaborazione predittiva, ciò che le parole e gli enunciati significano per un organismo è determinato in larga misura da ciò che quell'organismo si aspetta in risposta alle parole e agli enunciati — dagli stati sensoriali, motori, emozionali e cognitivi che il sistema predice a seguito dell'elaborazione linguistica. La parola «cane» attiva predizioni su suoni (il latrato), su sensazioni tattili (il pelo), su stati visivi (la forma del corpo), su sequenze motorie (il gesto di carezzare), su stati emotivi (l'affetto o il timore) — e il significato della parola, in questa prospettiva, non è una condizione di verità né una posizione in uno spazio vettoriale distribuzionale né un riferimento a un individuo, ma è questo complesso di predizioni multi-modali che l'elaborazione della parola attiva nel sistema cognitivo. Il significato è dunque integralmente **predittivo** e **incarnato**: è la struttura di attesa che la parola genera in un sistema che ha una storia di interazione sensomotoria con il mondo.

Questo resoconto del significato come struttura predittiva multi-modale affronta il problema del grounding che la semantica distribuzionale lasciava irrisolto. La semantica distribuzionale cattura le relazioni di similarità tra le distribuzioni di contesti di co-occorrenza, ma queste relazioni sono puramente linguistiche — chiuse all'interno del corpus testuale. Il paradigma predittivo fornisce un'architettura in cui le rappresentazioni linguistiche sono intrinsecamente connesse alle rappresentazioni percettive, motorie ed emozionali: una parola

significa qualcosa perché attiva predizioni su stati non linguistici dell'organismo che ha imparato la parola attraverso interazioni incarnate con il mondo. Il grounding non è un problema da risolvere aggiungendo un modulo percettivo al sistema linguistico: è incorporato nell'architettura stessa dell'elaborazione predittiva, perché le predizioni che la parola attiva si estendono naturalmente ai correlati percettivi e motori del suo referente. Il sistema non ha bisogno di una connessione esplicita tra le rappresentazioni linguistiche e quelle percettive: queste connessioni sono implicite nella struttura del modello generativo, appreso attraverso l'esposizione a stimoli linguistici e non linguistici in contesti in cui il linguaggio co-occorreva con esperienze sensorimotorie.

Andy Clark, nel volume *Surfing Uncertainty* del 2016, ha articolato la versione filosofica più elaborata del paradigma predittivo, estendendo il principio di energia libera di Friston in una proposta sulla natura della mente come sistema di anticipazione attiva. Nella prospettiva di Clark, il cervello non è un sistema che riceve input sensoriali e computa rappresentazioni del mondo: è un **sistema di generazione ipotetica** che produce continuamente predizioni sugli stati del mondo e confronta queste predizioni con i segnali sensoriali effettivi, utilizzando gli errori di predizione per affinare le ipotesi piuttosto che per costruire rappresentazioni dal basso. La percezione, in questo quadro, è essenzialmente allucinazione controllata: un processo di costruzione attiva di rappresentazioni guidata da predizioni — non da input — e corretta dagli errori. La correlazione tra l'allucinazione controllata della percezione ordinaria e l'allucinazione incontrollata degli stati psicotici non è soltanto una metafora: entrambi i fenomeni emergono dall'architettura predittiva, e la differenza tra percezione normale e psicosi è quantitativa, non qualitativa — è una differenza nel peso relativo attribuito alle predizioni top-down rispetto agli errori di predizione bottom-up.

Il collegamento con il programma dell'enattivismo esaminato nella sezione III.3 è di straordinario interesse filosofico e deve essere esplicitato. La fenomenologia enattivista di Varela, Thompson e Rosch aveva sostenuto che la cognizione non è computazione di rappresentazioni ma accoppiamento sensomotorio tra organismo e ambiente: il significato emerge dall'interazione dinamica, non da rappresentazioni interne. Il paradigma predittivo sembra, a prima vista, in tensione con questa posizione, perché parla esplicitamente di modelli interni, rappresentazioni e stati interni. Ma la tensione è più apparente che reale. Nell'inferenza attiva, l'azione è componente essenziale del processo di minimizzazione dell'energia libera: il sistema non si limita a costruire modelli passivi del mondo ma agisce sul mondo per renderlo conforme alle proprie predizioni. L'accoppiamento sensomotorio dell'enattivismo emerge come conseguenza dell'inferenza attiva: il sistema modifica l'ambiente per allineare le percezioni alle predizioni, e le predizioni sono calibrate dalle interazioni passate con l'ambiente, creando un circolo di mutua costituzione che l'enattivismo aveva identificato come struttura fondamentale della cognizione. La differenza è forse di accento piuttosto che di sostanza: l'enattivismo pone l'interazione al centro e tratta i modelli interni come strumenti dell'interazione; il paradigma predittivo pone il modello generativo al centro e tratta l'interazione come strategia di minimizzazione dell'energia libera. Ma entrambe le prospettive convergono

nell'affermare che la cognizione è irriducibile alla computazione di rappresentazioni su input sensoriali passivi.

La valutazione critica del paradigma predittivo deve affrontare almeno tre obiezioni significative. La prima è l'obiezione della **sovraunificazione**: il principio di energia libera è formulato con tale generalità che sembra capace di spiegare qualsiasi comportamento biologico, il che solleva il dubbio che sia empiricamente irrefutabile. Se qualsiasi sistema che persiste nel tempo «minimizza l'energia libera», allora il principio descrive una tautologia mascherata da spiegazione — ogni comportamento adattativo è, per definizione, riduttivo dell'energia libera, e la formulazione non aggiunge potere esplicativo. Friston risponde che le predizioni quantitative del modello — sui tempi di risposta, sulle traiettorie dell'apprendimento, sui pattern di attivazione neurale — sono verificabili e falsificabili, e che il valore del principio sta nella sua capacità di generare modelli specifici delle specifiche funzioni cognitive, non nella sua generalità. La seconda obiezione è quella del **cervello bayesiano ottimale**: il paradigma predittivo presuppone che il cervello implementi qualcosa di analogo all'inferenza bayesiana, ma l'inferenza bayesiana esatta è computazionalmente intrattabile per problemi di dimensionalità realistica. La risposta standard è che il cervello implementa un'**approssimazione variazionale** all'inferenza bayesiana — esattamente la distribuzione approssimata $q(x; \mu)$ nell'espressione dell'energia libera — e che questa approssimazione, sebbene non ottimale in senso bayesiano stretto, è computazionalmente efficiente e sufficientemente accurata per gli scopi adattativi. La terza obiezione concerne il **problema difficile della coscienza** che sarà esaminato nella sezione III.9: il paradigma predittivo offre un resoconto funzionale della cognizione come minimizzazione dell'energia libera, ma non spiega perché questa minimizzazione sia accompagnata da esperienza soggettiva. Come ogni approccio funzionale, rischia di lasciare irrisolto il problema difficile riducendo la coscienza alle sue funzioni senza spiegare la loro qualità fenomenica. Jakub Limanowski e Karl Friston hanno proposto che la coscienza stessa possa essere analizzata come un processo di inferenza attiva sul proprio corpo — la propriocezione come modello predittivo del sé — ma questa proposta, per quanto feconda, non scioglie il problema difficile ma lo sposta.

Per la questione del significato che guida questa monografia, il contributo più decisivo del paradigma predittivo è la sua capacità di fornire un'architettura cognitiva in cui il significato emerge come struttura di anticipazione radicata nell'esperienza incarnata. La parola «afferrare» non significa ciò che significa perché è definita in termini di altri termini (come nella semantica distribuzionale) o perché denota una proprietà (come nella semantica verocondizionale), ma perché attiva nel sistema cognitivo un insieme di predizioni sensomotorie — sulle posizioni della mano, sulla resistenza dell'oggetto, sulla sequenza di movimenti muscolari — che riflettono la storia di interazioni fisiche con oggetti afferrabili. Il significato è, in questo senso, il residuo predittivo dell'esperienza: la struttura di aspettative che l'interazione ripetuta con il mondo ha sedimentato nel sistema di generazione di ipotesi. Il grounding non è il collegamento di un sistema simbolico chiuso con il mondo esterno: è la struttura stessa del sistema cognitivo, che è organizzato come un generatore di predizioni calibrato sull'esperienza incarnata. Il paradigma predittivo fornisce così, almeno in abbozzo, la risposta

cognitiva al problema del grounding che la semantica distribuzionale e la semantica formale lasciavano irrisolto: non aggiungendo un modulo percettivo esterno al sistema linguistico, ma mostrando che il sistema linguistico è esso stesso un sistema predittivo profondamente intrecciato con le strutture percettive e motorie del cervello, attraverso un'architettura gerarchica in cui nessun livello è puramente linguistico o puramente sensoriale.

La sezione II.4bis ha lasciato aperto il problema dell'iperintensionalità come residuo irrisolto della semantica intensionale dei mondi possibili: nei contesti in cui la sostituzione di proposizioni logicamente equivalenti — vere negli stessi mondi possibili — non preserva il valore di verità, nessuna semantica che identifichi il contenuto proposizionale con un insieme di mondi possibili può distinguere tra le proposizioni che occorrono in tali contesti. La semantica bidimensionale di Chalmers mitiga il problema discriminando tra intensione primaria e secondaria, ma non lo dissolve: due necessità matematiche distinte — «7 è primo» e «ogni numero ha un successore» — condividono la stessa intensione primaria (entrambe sono vere in ogni mondo centrato possibile), e tuttavia non sono intercambiabili nel contesto «Mario crede che___». La sostituzione di proposizioni necessarie vere in contesti doxastici, epistemici, esplicativi o controfattuali rivela una struttura semantica più fine di qualsiasi funzione dai mondi possibili ai valori di verità.

Il programma della **semantica dei fattori di verità**(*truthmaker semantics*), sviluppato nella sua forma sistematica da Kit Fine in una serie di scritti tra il 2012 e il 2017 — «A Difficulty for the Possible Worlds Analysis of Counterfactuals» (2012), «Truthmaker Semantics for Intuitionistic Logic» (2014), «Truthmaker Semantics» (2017) — non prende le mosse da considerazioni linguistiche o da problemi di semantica formale in senso stretto, ma da una questione metafisica che la tradizione angloamericana aveva posto con crescente urgenza a partire dagli anni Novanta: quella dei **fattori di verità** (*truthmakers*), gli stati del mondo in virtù dei quali le proposizioni sono vere. La domanda classica dei fattori di verità è: se la proposizione «il gatto è sul tappeto» è vera, che cosa nel mondo la rende vera? La risposta ovvia — che la rende vera il fatto che il gatto sia sul tappeto — è filosofica mente più densa di quanto appaia in superficie, perché introduce la nozione di **stato di cose** (*state of affairs*) come entità ontologica dotata di struttura: non il semplice gatto né il semplice tappeto, ma la relazione di trovarsi-su che li connette in questo momento, come questo particulare configurazione della realtà.

La versione della semantica dei fattori di verità che Fine ha elaborato adotta un'ontologia di **stati** (*states*) — entità parziali che rappresentano frammenti del mondo, non totalità complete come i mondi possibili. La differenza strutturale tra stati e mondi possibili è fondamentale: un mondo possibile è un'entità massimale che determina il valore di verità di ogni proposizione; uno stato è un'entità parziale che può determinare il valore di verità di alcune proposizioni senza determinarlo per tutte. Lo stato che consiste nel gatto essendo sul tappeto verifica «il gatto è sul tappeto» senza nulla dire di «il cane abbaia» o di «2+2=4». La parzialità è la caratteristica costitutiva degli stati: essi abitano uno spazio ontologico di dimensione inferiore ai mondi possibili, e possono essere costruiti, combinati e confrontati in modi che i mondi possibili non ammettono.

La struttura algebrica degli stati presuppone una relazione di **fusione** (*fusion*): dati due stati s_1 e s_2, esiste uno stato $s_1 \sqcup s_2$ che è la loro fusione — lo stato che include tutto ciò che include s_1 e tutto ciò che include s_2. La fusione è commutativa ($s_1 \sqcup s_2 = s_2 \sqcup s_1$), associativa ($(s_1 \sqcup s_2) \sqcup s_3 = s_1 \sqcup (s_2 \sqcup s_3)$) e idempotente ($s \sqcup s = s$). Questi assiomi fanno dello spazio degli stati un **semireticolo superiore**: un insieme ordinato parzialmente in cui ogni coppia di elementi ha un limite superiore (la fusione), ma non necessariamente un limite inferiore. L'ordinamento parziale $\leq$ è definito da: $s_1 \leq s_2$ se e soltanto se $s_1 \sqcup s_2 = s_2$ — cioè s_1 è una **parte** di s_2 nel senso mereologico. Esiste uno stato minimale 0 (lo stato vuoto, privo di contenuto), e la fusione generalizzata di una famiglia di stati è il loro join nel semireticolo.

La relazione semantica centrale nella semantica dei fattori di verità non è la verità di una proposizione in un mondo ma la relazione binaria di **verifica** (*verification*) tra uno stato e una proposizione: $s \Vdash P$, letto «lo stato s verifica la proposizione P». Le condizioni di verifica per i connettivi proposizionali sono le seguenti. La verifica della **congiunzione**:

$s \Vdash P \wedge Q$ sse $\exists s_1, s_2$: $s_1 \Vdash P$ e $s_2 \Vdash Q$ e $s = s_1 \sqcup s_2$

Lo stato che verifica una congiunzione è la fusione di due stati, uno dei quali verifica il primo congiunto e l'altro il secondo. La richiesta che lo stato sia esattamente la fusione — e non un superstato arbitrario che la contiene — è la condizione di **rilevanza esatta** (*exact truthmaking*): lo stato deve essere esattamente proporzionato alla proposizione che verifica, senza includere informazione irrilevante. La verifica della **disgiunzione**:

$s \Vdash P \vee Q$ sse $s \Vdash P$ o $s \Vdash Q$

Uno stato verifica una disgiunzione se verifica uno dei disgiunti — non la loro fusione, ma l'uno o l'altro. La verifica dell'**implicazione** richiede un trattamento più sofisticato che incorpora la relazione di **falsificazione**:

$s \Vdash P \rightarrow Q$ sse $\forall s'$: se $s' \Vdash P$ allora $s \sqcup s' \Vdash Q$

La implicazione è verificata dallo stato s se, per ogni stato che verifica P, la fusione di s con quello stato verifica Q. Questa è la condizione di **rilevanza** dell'implicazione: lo stato s deve contribuire alla verifica di Q in modo dipendente dalla verifica di P, non verificare Q indipendentemente. Infine, la verifica della **negazione** richiede la nozione complementare di **falsificazione**: uno stato s **falsifica** P (si scrive $s \dashv P$) se s è incompatibile con ogni stato che verifica P — se nessuna fusione di s con uno stato verificante P è uno stato coerente. La negazione è verificata quando viene falsificato il contenuto negato:

$s \Vdash \neg P$ sse $s \dashv P$

La condizione di incompatibilità è definita in termini di uno spazio di stati *coerenti*: un insieme di stati è coerente se esiste un mondo possibile che «include» tutti gli stati nell'insieme (nel senso tecnico che ogni stato nell'insieme è una parte del mondo). Due stati s_1 e s_2 sono **incompatibili** se $\{s_1, s_2\}$ non è coerente — se nessun mondo li include entrambi. Una proposizione P è **veridica** in un mondo

w se w include qualche stato s tale che s ⊩ P. La verità in un mondo è dunque derivata dalla relazione di verifica degli stati, non primitiva.

Il risultato che rende la semantica dei fattori di verità strumentalmente feconda per il problema dell'iperintensionalità è il seguente. Due proposizioni P e Q sono **fattori di verità realmente equivalenti** se e soltanto se ogni stato che verifica P verifica Q e viceversa. Due proposizioni possono condividere gli stessi mondi possibili — essere logicamente equivalenti — pur differendo nei loro fattori di verità. Le proposizioni «7 è primo» e «7 è primo e (7 è primo o 7 non è primo)» sono logicamente equivalenti: vere negli stessi mondi. Ma i loro fattori di verità sono diversi: uno stato che verifica la prima è uno stato che riguarda il numero 7 e la sua relazione con i propri divisori; uno stato che verifica la seconda è la fusione di quello stato con uno stato che verifica la tautologia «7 è primo o 7 non è primo» — uno stato ulteriore, irrilevante alla materia principale. La tautologia aggiuntiva è semanticamente visibile nella struttura dei fattori di verità anche se è invisibile nella struttura dei mondi possibili.

Fine ha elaborato, su questa base, una nozione di **implicazione rilevante** (*relevant entailment*) che cattura la struttura logica interna alle proposizioni senza dipendere dalla nozione di mondo possibile. Una proposizione P implica rilevantemente Q (P ⊩_r Q) se e soltanto se ogni fattore di verità esatto di P include un fattore di verità esatto di Q — se ogni stato che verifica P include uno stato che verifica Q. L'implicazione rilevante non è semplicemente contenimento dei modelli: è contenimento dei fattori di verità. La congiunzione P ∧ Q implica rilevantemente P, perché ogni stato $s_1 \sqcup s_2$ che verifica P ∧ Q include lo stato s_1 che verifica P. La tautologia P ∨ ¬P non implica rilevantemente nessuna proposizione contingente: i suoi fattori di verità — stati che verificano P oppure stati che verificano ¬P — non includono necessariamente nessuno stato che verifichi, diciamo, «il sole è una stella». Questo è il risultato che la logica rilevante cercava di catturare dall'interno di un sistema puramente sintattico: la semantica dei fattori di verità si fonda su una base ontologica indipendente, senza richiedere la revisione del principio del terzo escluso né la rinuncia alla logica classica a livello dei mondi possibili.

La connessione con il problema dei **contesti iperintensionali** si articola nel modo seguente. Fine propone che i contesti doxastici, epistemici ed esplicativi siano sensibili non alle proposizioni intensionali (insiemi di mondi) ma ai **contenuti fattori di verità** — alle strutture di verifica delle proposizioni. Quando si chiede «perché P?» e si risponde «perché Q», la risposta è appropriata soltanto se i fattori di verità di Q sono in qualche modo rilevanti per i fattori di verità di P — se c'è una relazione di contribuzione causale o costitutiva tra gli stati che verificano Q e quelli che verificano P. La spiegazione non è preservata dalla sostituzione di proposizioni logicamente equivalenti: «perché la somma degli angoli interni di un triangolo è 180°?» ha come risposta appropriata «perché la geometria dello spazio è euclidea», non «perché è il caso che (la somma degli angoli è 180° e la luna non è fatta di formaggio verde) o (la somma degli angoli non è 180°)», sebbene le due proposizioni siano logicamente equivalenti. Il contenuto esplicativo dipende dai fattori di verità — dagli stati che verificano la proposizione esplicante — e questi differiscono radicalmente tra le due formulazioni equivalenti.

Fine ha applicato questo quadro alla semantica dei **controfattuali** in modo che supera alcune difficoltà note dell'analisi di Lewis basata sulla similarità tra mondi possibili. Un controfattuale «se P fosse il caso, Q sarebbe il caso» (P □→ Q) è vero, nell'analisi lewisiana, se i mondi possibili più simili al mondo attuale in cui P è vera sono mondi in cui anche Q è vera. Ma la nozione di similarità tra mondi è notoriamente vaga e contestata: Lewis stesso ammette che la metrica di similarità dipende dal contesto e non può essere resa completamente precisa senza perdere la generalità dell'analisi. La semantica dei fattori di veritàoffre un'alternativa: P □→ Q è vero se ogni stato che *giustifica* P (nel senso di verificarlo in modo minimale, senza elementi irrilevanti) si combina con il contesto fattuale in modo da giustificare Q. La nozione di giustificazione minimale rimpiazza quella di mondi più simili, e la valutazione del controfattuale non richiede di confrontare mondi interi ma di esaminare gli stati parziali rilevanti per P e per Q nel contesto fattuale dato.

La dimensione filosofica più ampia del programma di Fine riguarda la nozione di **contenuto esatto** (*exact content*) di una proposizione: il complesso dei suoi fattori di verità esatti, che specifica non soltanto in quali mondi la proposizione è vera ma per quale ragione vi è vera, attraverso quali stati del mondo. Il contenuto esatto è più fine dell'intensione, distinguendo proposizioni che nessuna semantica intensionale può differenziare. Ma è anche strutturato in modo diverso dalla struttura sintattica della proposizione: due formulazioni sintatticamente diverse possono avere lo stesso contenuto esatto (se hanno gli stessi fattori di verità), e due formulazioni sintatticamente identiche possono avere contenuti esatti diversi in contesti di interpretazione diversi. Il contenuto esatto è dunque un terzo livello di struttura semantica — distinto sia dalla struttura sintattica sia dall'intensione — che cattura la rilevanza ontologica della proposizione: il frammento del mondo che la rende vera.

Questa nozione si connette direttamente, e non casualmente, alle problematiche della Parte Prima. Il problema dell'ontologia matematica — che cosa rende veri i teoremi matematici — riceve una formulazione precisa nella semantica dei fattori di verità: i fattori di verità dei teoremi matematici sono stati di cose matematiche, e la domanda di quale sia la loro natura ontologica è la domanda dello strutturalismo matematico posta in termini fattori di verità. L'Assioma di Univalenza della HoTT, che identifica l'uguaglianza con l'equivalenza omotopica, ha una formulazione a fattori di verità naturale: i fattori di verità dell'identità A = B sono le equivalenze omotopiche tra A e B, non alcunché di più sottile o più primitivo. Il predicativismo, a sua volta, può essere formulato come una restrizione sui fattori di verità ammissibili: le proposizioni matematiche predicativamente legittime sono quelle i cui fattori di verità non presuppongono come dato il dominio totale su cui quantificano. La semantica dei fattori di verità fornisce dunque non soltanto un'alternativa alla semantica intensionale per i problemi del linguaggio naturale, ma un quadro metafisico che illumina le questioni ontologiche della matematica con strumenti più fini di quelli disponibili nell'approccio estensionale o intensionale.

Il confronto con la logica della **rilevanza** (*relevance logic*) — sviluppata da Wilhelm Ackermann, Alan Ross Anderson e Nuel Belnap a partire dagli anni Cinquanta come alternativa alla logica classica che elimina le paradossalità dell'implicazione

materiale — chiarisce la posizione della semantica dei fattori di verità nel panorama delle logiche non classiche. La logica della rilevanza aveva tentato di catturare l'intuizione che in una valida implicazione «A implica B», A e B devono condividere variabili — deve esserci una connessione di contenuto tra antecedente e conseguente. Il risultato era un sistema assiomatico potente ma dal fondamento semantico problematico: le semantiche dei «mondi impossibili» usate per le logiche della rilevanza sono tecnicamente sofisticate ma ontologicamente oscure. La semantica dei fattori di verità di Fine raggiunge lo stesso obiettivo — catturare la rilevanza — attraverso una semantica che è fondata sull'ontologia degli stati parziali: la rilevanza emerge dalla struttura dei fattori di verità, non da assiomi sintattici sulla condivisione di variabili. Il risultato è che le leggi della logica della rilevanza emergono come teoremi della semantica dei fattori di verità, e le loro eccezioni sono precisamente i principi che la semantica dei fattori di verità esclude per ragioni ontologiche chiare. La semantica dei fattori di verità non è dunque soltanto un'alternativa tecnica alla semantica dei mondi possibili: è una riconfigurazione profonda del rapporto tra logica, semantica e ontologia, che affronta il problema dell'iperintensionalità non aggiungendo pezzi al quadro dei mondi possibili ma costruendo un quadro ontologicamente più ricco in cui la struttura del mondo — la distribuzione degli stati parziali — determina la struttura del significato in modo più fine di quanto nessuna semantica precedente avesse potuto fare.

III.9 — Il problema difficile della coscienza e la questione del significato soggettivo

L'ultima questione che la Parte Terza deve affrontare è quella che connette più direttamente le scienze cognitive alla filosofia del linguaggio e alla Parte Quarta: il problema difficile della coscienza, formulato da David Chalmers in «Facing Up to the Problem of Consciousness» (1995) con una distinzione che ha ridefinito il panorama della filosofia della mente. I **problemi facili** della coscienza — facili non nel senso di tecnicamente semplici, ma nel senso di concettualmente non misteriosi una volta che si è compreso il tipo di spiegazione richiesta — comprendono la spiegazione di funzioni cognitive come la capacità di discriminare stimoli e reagire in modo differenziato, l'integrazione delle informazioni, la capacità di effettuare un report degli stati mentali, l'accesso alle proprie elaborazioni interne, il controllo del comportamento da parte degli stati cognitivi. Per spiegare questi fenomeni è sufficiente — in linea di principio — identificare i meccanismi neurali o funzionali che li realizzano: si tratta di problemi empirici difficili, non di misteri filosófici.

Il **problema difficile** (*hard problem*) è diverso in natura: è il problema di spiegare **perché** i processi neurali o funzionali che realizzano le funzioni cognitive sono accompagnati da esperienza soggettiva — perché c'è qualcosa che si prova nel vedere il rosso, nel sentire dolore, nell'ascoltare una melodia. Questa dimensione qualitativa dell'esperienza — i **qualia** — non sembra catturabile da nessuna descrizione funzionale o fisica: si può descrivere integralmente la neurofisiologia della percezione del rosso, specificare le funzioni computazionali che realizza, e

tuttavia qualcosa sembra restare fuori dalla descrizione — il fatto che vedere il rosso abbia la qualità che ha, che «sembri rosso» dall'interno.

Thomas Nagel ha formulato questo punto con il celebre esperimento mentale di «What Is It Like to Be a Bat?» (1974). I pipistrelli percepiscono il mondo principalmente attraverso l'ecolocalizzazione — un sistema sensoriale radicalmente diverso dalla visione. C'è qualcosa che è come essere un pipistrello — qualcosa che è come percepire il mondo attraverso l'ecolocalizzazione — ma non possiamo sapere cosa sia, perché le nostre facoltà di immaginazione sono calibrate sulla nostra esperienza sensoriale e non possono essere proiettate in modo affidabile sull'esperienza sensoriale del pipistrello. Il punto filosofico di Nagel non è soltanto che l'ecolocalizzazione è difficile da immaginare per noi, ma che l'esperienza soggettiva ha un carattere essenzialmente prospettico — è sempre esperienza di un soggetto determinato, vista «dall'interno» — che le descrizioni oggettive in terza persona non possono catturare per principio.

L'argomento di **Mary la scienziata** di Frank Jackson (1982) è la versione più precisa di questo punto. Mary è una scienziata che ha vissuto tutta la vita in una stanza in bianco e nero, senza mai vedere colori. Tuttavia conosce tutti i fatti fisici sulla percezione del colore: conosce la struttura della retina, i meccanismi di trasduzione dei coni, le vie neurali, le aree corticali coinvolte, le descrizioni funzionali dei processi percettivi. Quando esce dalla stanza e vede il rosso per la prima volta, impara qualcosa di nuovo? Jackson sostiene che sì: impara come appare il rosso, cosa si prova a vedere il rosso — un fatto che non era contenuto nella sua enciclopedica conoscenza fisica. Se impara qualcosa di nuovo, c'è un fatto sulla sua esperienza di vedere il rosso che non era catturato dalla conoscenza fisica completa; dunque esistono fatti non fisici sull'esperienza cosciente, e il fisicalismo è falso.

Le risposte fisicaliste all'argomento di Mary sono numerose. La risposta di **Dennett** è che Mary non impara nulla di genuinamente nuovo: se veramente conosceva tutti i fatti fisici, conosceva già come appare il rosso, nel senso di avere tutte le disposizioni comportamentali e neurali appropriate; ciò che esperisce alla prima visione del rosso è l'attivazione di disposizioni già presenti, non l'acquisizione di un fatto nuovo. La risposta dell'**abilità** (*ability reply*) di Lewis e Nemirow è che Mary non acquisisce nuova conoscenza proposizionale ma nuove abilità: sa come si percepisce il rosso nel senso di poterlo immaginare, riconoscere e ricordare — ma si tratta di abilità pratiche, non di conoscenza di nuovi fatti. La risposta del **concetto fenomenico** (*phenomenal concept strategy*) di Papineau e Loar è che Mary acquisisce un nuovo concetto — il concetto fenomenico di rosso — che si riferisce allo stesso fatto fisico di cui già disponeva, ma attraverso una modalità di concezione diversa: la cognizione diretta, in prima persona, dell'esperienza stessa.

La **concezione zombie** di Chalmers porta il problema alla sua forma più netta. Uno **zombie filosofico** è un essere fisicamente identico a noi — con gli stessi neuroni, gli stessi processi funzionali, lo stesso comportamento — ma privo di esperienza cosciente: non c'è nulla che sia come essere uno zombie. Chalmers sostiene che gli zombie sono **metafisicamente possibili**: non c'è contraddizione logica nel concepire un essere fisicamente identico a noi ma privo di qualia. Se

sono metafisicamente possibili, allora la coscienza non è logicamente superveniante sulla fisica — non è necessariamente determinata dai fatti fisici — e il fisicalismo di tipo-A (riduzionismo immediato) è falso. Il fisicalista risponde o negando la possibilità metafisica degli zombie (tipo-B: la supervenianza è necessaria anche se non concettualmente ovvia) o accettando un dualismo di proprietà (tipo-D: la coscienza è una proprietà fondamentale non riducibile).

La connessione con il problema del significato è diretta. John Searle ha sostenuto, nell'argomento della Stanza Cinese che si esaminerà nella sezione IV.2, che la semantica — il contenuto genuino — richiede la coscienza: soltanto un sistema cosciente può avere intenzionalità intrinseca, ovvero può avere stati che si riferiscono genuinamente a qualcosa. I sistemi artificiali hanno intenzionalità derivata — si riferiscono a oggetti nel mondo soltanto nella misura in cui gli utenti o i progettisti attribuiscono loro un significato — ma non intenzionalità originaria. Se Searle ha ragione, il problema del significato nei sistemi artificiali non è separabile dal problema della coscienza: un sistema che non ha esperienza soggettiva non può avere contenuto semantico genuino, indipendentemente dalla complessità del suo comportamento linguistico.

Questa tesi non è pacificamente accettata. Dennett nega che l'intenzionalità intrinseca — nel senso di un contenuto semantico non derivato da interpretazione — esista anche negli esseri umani: ogni attribuzione di significato è prospettica, è il prodotto dell'adozione dello **stance intenzionale** (*intentional stance*) — la strategia interpretativa che attribuisce credenze, desideri e razionalità a un sistema per predirne il comportamento. Non vi è differenza di principio tra attribuire intenzionalità a un essere umano e attribuirla a un sistema artificiale: entrambe le attribuzioni sono strategie predittive, non descrizioni di proprietà intrinseche. La risposta di Searle è che questa dissoluzione pragmatica manca il punto filosofico: c'è qualcosa che è come capire il cinese per un parlante nativo del cinese, e questo qualcosa è assente nello zombie linguistico che manipola simboli senza comprenderli. La questione — come ogni questione autenticamente filosofica esaminata in questa monografia — non si chiude nel confronto delle posizioni ma si approfondisce: e il suo approfondimento richiede di esaminare i sistemi artificiali di elaborazione del linguaggio, ai quali si volge la Parte Quarta.

III.10 — La mente estesa e la cognizione 4E: il significato oltre i confini del cranio

La concezione standard della cognizione — sia nella versione computazionale classica sia nella versione connessionista — localizza i processi mentali all'interno del cervello o, al più, all'interno del sistema nervoso. Il significato, in questa concezione, è un prodotto interno: emerge dall'attività di sistemi neurali che elaborano rappresentazioni, e la sua connessione con il mondo esterno passa attraverso i canali sensoriali in ingresso e motori in uscita. Un programma di ricerca che ha acquisito rilevanza crescente a partire dagli anni Novanta ha messo in questione questa localizzazione, proponendo che i processi cognitivi — inclusa la formazione e il mantenimento dei contenuti significativi — non siano confinati

nel cranio ma si estendano nell'ambiente e negli artefatti cognitivi con cui l'agente interagisce.

Andy Clark e David Chalmers hanno articolato la versione filosofica più influente di questa proposta nell'articolo «The Extended Mind» (1998), che ha innescato un dibattito che continua tutt'ora. L'argomento centrale è il **Principio di Parità** (*Parity Principle*): se in certi compiti cognitivi il mondo gioca lo stesso ruolo funzionale che giocherebbe un processo interno al cranio, allora quell'aspetto del mondo è legittimamente parte della mente. Il principio non è una tesi empirica sulla struttura del cervello ma una tesi filosofica sulla delimitazione dei confini della mente: esso chiede che il criterio di appartenenza alla mente sia il ruolo funzionale del processo, non il suo substrato materiale o la sua localizzazione spaziale.

L'esperimento mentale fondamentale è il confronto tra Inga e Otto. Inga è una persona con memoria intatta che vuole visitare il MoMA a New York: ricorda che il museo si trova sulla Cinquantatreesima Strada, usa questa credenza per pianificare il percorso. Otto è una persona con deficit di memoria dovuto all'Alzheimer che porta con sé un taccuino in cui ha scritto informazioni importanti: vuole visitare il MoMA, consulta il taccuino, trova l'indirizzo, usa questa informazione per pianificare il percorso. Clark e Chalmers sostengono che il taccuino di Otto svolge esattamente la stessa funzione cognitiva che la memoria di Inga svolge per Inga: fornisce l'accesso all'informazione rilevante quando è necessaria, guida la pianificazione e l'azione nella stessa struttura causale. Se ammettiamo che la credenza di Inga — localizzata nel cervello — sia una credenza genuina, allora per il Principio di Parità dobbiamo ammettere che anche il contenuto del taccuino di Otto costituisce una credenza — una credenza estesa nell'ambiente. La mente di Otto si estende nel taccuino.

Le condizioni che Clark e Chalmers pongono perché un'informazione esterna conti come stato cognitivo esteso sono tre. L'informazione deve essere **prontamente disponibile** — accessibile quando rilevante per il compito in corso, senza richiedere un processo di ricerca esterna lunga e laboriosa. Deve essere **automaticamente asserita** — accettata dal sistema cognitivo senza ulteriori processi di verifica critica, così come le credenze interne vengono usate senza essere messe in discussione ad ogni utilizzo. E il suo contenuto deve essere **derivato dall'endorsement passato** — basato sull'aver previamente inserito e validato l'informazione nel sistema esterno, così come le credenze interne sono basate su processi di formazione precedenti. Queste condizioni escludono gli elenchi telefonici o i motori di ricerca dall'essere parti della mente cognitiva: essi non soddisfano la condizione di endorsement automatico perché le informazioni che contengono non sono state validate dal soggetto e non vengono accettate senza verifica critica.

La tesi della mente estesa ha implicazioni dirette per il problema del grounding semantico esaminato nelle sezioni precedenti. Se le rappresentazioni mentali si estendono nell'ambiente, allora la connessione tra il simbolo e il mondo non passa soltanto attraverso i processi interni di elaborazione ma attraverso le pratiche d'uso degli artefatti cognitivi in contesti d'azione. Il taccuino di Otto non è un serbatoio di simboli non interpretati: è uno strumento che Otto ha imparato

a usare in pratiche d'azione determinate, e il suo significato è costituito dall'intreccio tra le marcature sul taccuino, le pratiche d'uso che Otto ha sviluppato, e le relazioni causali con gli oggetti del mondo che quelle pratiche codificano. Il grounding, in questa prospettiva, non è un problema da risolvere attraverso la connessione di un sistema simbolico con un sistema percettivo; è già risolto nelle pratiche incarnate e situate di uso degli artefatti cognitivi.

Edwin Hutchins ha sviluppato, in *Cognition in the Wild* (1995), una teoria della **cognizione distribuita** che radicalizza la proposta di Clark e Chalmers estendendo la cognizione non soltanto agli artefatti ma alle pratiche sociali e organizzative. Hutchins studia la navigazione delle navi da guerra della marina statunitense come sistema cognitivo distribuito: il calcolo della posizione della nave, la pianificazione della rotta, l'identificazione dei punti di riferimento non sono operazioni compiute da un singolo individuo ma da un sistema di individui che interagiscono con strumenti, con grafici, con procedure istituzionalizzate e con ciascun altro. La rappresentazione rilevante — la posizione della nave, il corso da seguire — non è localizzata in nessuna mente individuale ma è distribuita attraverso il sistema sociotecnico nella sua interezza.

La cognizione distribuita modifica profondamente il quadro del significato: i concetti non sono proprietà di menti individuali ma risorse condivise di pratiche collettive. Il significato di «rotta» nell'ambiente della navigazione navale non è determinato da rappresentazioni interne al cervello del comandante ma dalle pratiche condivise di costruzione della rotta, lettura delle carte, calibrazione degli strumenti, coordinazione tra i membri dell'equipaggio. Il grounding del concetto «rotta» è distribuito attraverso una rete di artefatti, pratiche e agenti che costituiscono insieme il sistema cognitivo della navigazione. Questa prospettiva entra in un dialogo productivo con l'inferenzialismo di Brandom: le pratiche inferenziali che costituiscono il significato non sono pratiche di individui isolati ma pratiche di comunità discorsive e pratiche sociotecniche.

La critica più incisiva alla mente estesa è stata formulata da Frederick Adams e Kenneth Aizawa nel volume *The Bounds of Cognition* (2008). Adams e Aizawa sostengono che il Principio di Parità è troppo liberale e porterebbe a conseguenze controverse: dovremmo concludere che il martello è parte della mente del falegname durante il lavoro, o che la penna è parte della mente del narratore durante la scrittura. La loro risposta è che la cognizione richiede non soltanto il ruolo funzionale ma anche la presenza di **rappresentazioni non derivate** (*non-derived representations*): stati interni la cui proprietà semantica è intrinseca e non dipende dall'interpretazione di agenti esterni. Il taccuino di Otto non ha rappresentazioni non derivate — le marcature significano ciò che significano soltanto perché Otto le interpreta così. Il cervello di Inga ha invece stati con contenuto intrinseco che non dipende dall'interpretazione esterna. Clark ha risposto che la distinzione tra contenuto intrinseco e derivato presuppone ciò che è in discussione: il problema del grounding si applica allo stesso modo ai neuroni e agli artefatti, e non vi è ragione di principio per cui i processi neurali abbiano contenuto intrinseco mentre i processi negli artefatti lo abbiano soltanto derivatamente.

Il dibattito tra Clark-Chalmers e Adams-Aizawa riproduce, al livello della cognizione incarnata e distribuita, l'aporia centrale di questa monografia: il problema di dove risieda il significato, se nella struttura interna del sistema o nella relazione tra il sistema e il mondo. La mente estesa radicalizza la risposta relazionale: il significato non è né dentro il cranio né negli oggetti del mondo, ma nelle pratiche d'uso che connettono agenti, artefatti e mondo in sistemi funzionalmente integrati. Questa radicalizzazione non risolve il problema del grounding ma lo riconfigura: invece di chiedersi come i simboli interni al cervello acquisiscano contenuto, si chiede come le pratiche socio-tecniche costituiscano significato nelle interazioni tra agenti e ambienti.

PARTE QUARTA: INTELLIGENZA ARTIFICIALE, SIMBOLI E LA QUESTIONE DEL SIGNIFICATO MECCANICO

IV.1 — Il test di Turing: comportamento, imitazione e la struttura del problema dell'altra mente

Il punto di partenza della filosofia dell'intelligenza artificiale è un articolo di Alan Turing pubblicato nel 1950 su *Mind* con il titolo «Computing Machinery and Intelligence», che ha il merito straordinario di porre il problema nella sua forma più operazionalmente precisa e filosofica mente più sfuggente al tempo stesso. Turing non propone una definizione di intelligenza né postula una teoria dell'intenzionalità; propone invece di sostituire la domanda «le macchine possono pensare?» — che ritiene insolubile a causa dell'indeterminatezza dei termini coinvolti — con una domanda operazionalmente definita: una macchina può imitare un essere umano in un gioco di interrogazione? Il **gioco dell'imitazione** (*Imitation Game*) nelle sue versioni originali e nella versione che è diventata nota come Test di Turing prevede che un interrogante comunichi attraverso un canale testuale con due interlocutori, uno dei quali è un essere umano e l'altro una macchina, e cerchi di distinguerli. Se la macchina riesce a ingannare l'interrogante nel trenta per cento dei casi o più nel corso di un interrogatorio di cinque minuti, la supera.

L'argomento di Turing in favore dell'idoneità di questo test è che le obiezioni all'intelligenza delle macchine non reggono sotto esame. L'obiezione della coscienza — soltanto un sistema cosciente può genuinamente pensare — è irrisolvibile perché la coscienza degli altri non è direttamente accessibile: non siamo in grado di dimostrare che gli altri esseri umani siano coscienti piuttosto che zombie filosófici nel senso di Chalmers. Se ammettiamo che gli altri umani pensino sulla base del loro comportamento, allora per coerenza dobbiamo riconoscere la medesima inferenza per le macchine il cui comportamento è indistinguibile. L'obiezione teologica, l'obiezione dell'originalità (le macchine fanno soltanto ciò per cui sono programmate), l'obiezione del continuo di Lady Lovelace (citata dallo stesso Turing) — tutte vengono esaminate e respinte. Il

nucleo della risposta di Turing è comportamentista nel metodo: ciò che conta per l'attribuzione dell'intelligenza è il comportamento osservabile, non il substrato che lo produce.

La difficoltà filosofica del test di Turing non è tecnica ma concettuale: esso risponde alla domanda «questa macchina si comporta come se comprendesse?» e non alla domanda «questa macchina comprende?». Questa distinzione non è netta soltanto per il filosofo: è netta anche per il programma ordinario di linguistica cognitiva e di psicologia, che mira a inferire strutture interne — rappresentazioni, processi, capacità — dal comportamento osservato, e che sa bene che comportamento e struttura interna non si identificano. Un sistema può imitare un comportamento intelligente attraverso meccanismi che non hanno nulla in comune con i meccanismi che producono il comportamento intelligente nell'essere umano: il risultato osservato può essere identico mentre la struttura causale differisce radicalmente. Questa possibilità — denominata da Ned Block la distinzione tra **intelligenza-S** (intelligenza come comportamento sistematicamente adeguato) e **intelligenza-P** (intelligenza come capacità di produrre comportamento adeguato attraverso i meccanismi appropriati) — mostra che il test di Turing risponde alla prima domanda ma non alla seconda.

Il problema dell'altra mente è, in questa luce, strutturalmente diverso per gli esseri umani e per i sistemi artificiali. Per gli esseri umani, l'inferenza all'esistenza di menti negli altri poggia su due pilastri: l'analogia per mezzo dell'identità di substrato (gli altri umani hanno un sistema nervoso simile al mio, e il mio sistema nervoso produce esperienza cosciente, dunque è probabile che i loro la producano) e la teoria evolutiva della continuità (la mente umana è il prodotto di un processo evolutivo che la ha generata in tutti i membri della specie). Per i sistemi artificiali, il primo pilastro è assente — il substrato è radicalmente diverso — e il secondo è irrilevante. L'analogia funzionale soltanto — il comportamento simile — è una base di inferenza molto più debole. La questione di se i sistemi artificiali abbiano esperienza cosciente, stati intenzionali genuini e accesso al significato non si risolve con l'esame del comportamento esterno, ma richiede o una teoria della coscienza sufficientemente fondamentale da stabilire quando le sue condizioni sono soddisfatte (Chalmers, Tononi) o un argomento diretto sulle strutture interne del sistema (Searle).

La configurazione teorica in cui Turing inserisce la sua proposta è quella della **computazione universale**: una macchina di Turing universale può simulare qualsiasi altra macchina di Turing, e dunque qualsiasi processo computazionale discreto. Se la cognizione è computazione — se i processi mentali sono processi computazionali su rappresentazioni simboliche, come la LOTH di Fodor afferma — allora una macchina di Turing sufficientemente potente può in principio simulare qualsiasi processo cognitivo. La chiesa-Turing tesi afferma che ogni funzione calcolabile da un procedimento effettivo è calcolabile da una macchina di Turing, il che trasforma la tesi computazionalista in una tesi sulla universalità della macchina di Turing rispetto ai processi cognitivi. È questa tesi — non il semplice test comportamentale — che costituisce il nucleo teorico dell'ipotesi dell'intelligenza artificiale forte (*Strong AI*).

IV.2 — La Stanza Cinese: intenzionalità intrinseca, semantica e i limiti dell'argomentazione sintattica

L'argomento filosofico più influente contro l'intelligenza artificiale forte — contro la tesi che un sistema computazionale che simula con sufficiente accuratezza il comportamento cognitivo umano ipso facto possieda stati mentali genuini — è la **Stanza Cinese** di John Searle, presentata per la prima volta in «Minds, Brains, and Programs» (1980, *Behavioral and Brain Sciences*) e sviluppata poi in *Intentionality: An Essay in the Philosophy of Mind* (1983) e in *The Rediscovery of the Mind* (1992).

L'esperimento mentale è costruito con cura argomentativa notevole. Si immagini che Searle stesso — che non capisce il cinese — venga rinchiuso in una stanza con un ampio insieme di simboli cinesi, con regole per manipolare i simboli scritte in inglese (che capisce) e con un meccanismo per ricevere e inviare sequenze di simboli all'esterno. Dall'esterno, madrelingua cinesi inviano domande scritte in cinese; Searle, seguendo le regole, manipola i simboli e produce risposte in cinese che risultano appropriate, tali da ingannare i parlanti cinesi nel ritenere di comunicare con un madrelingua. Dal punto di vista esterno, la stanza supera il test di Turing per il cinese. Ma Searle, all'interno, non capisce una sola parola di cinese: manipola simboli formali secondo regole formali, senza alcuna comprensione del contenuto semantico.

L'argomento si applica a qualsiasi sistema di intelligenza artificiale classica che simula la comprensione del linguaggio attraverso la manipolazione di simboli secondo regole: tale sistema ha **sintassi** — struttura formale che determina quali operazioni sono applicabili a quali simboli — ma non **semantica** — contenuto, significato, riferimento al mondo. La manipolazione sintattica, per quanto sofisticata, non produce semantica: questa è la tesi principale di Searle, denominata il **principio fondamentale** della sua argomentazione. I formalismo sintattico è il terzo incomodo tra la struttura e il contenuto: può mediare tra i due (le regole sintattiche fissano le condizioni di derivazione che preservano le relazioni semantiche) ma non può generare il contenuto dal nulla.

La risposta più diffusa all'argomento — che Searle chiama la **risposta del sistema** — afferma che la confusione sta nel guardare soltanto alla parte del sistema (Searle all'interno della stanza) anziché al sistema nella sua interezza. Searle non capisce il cinese, ma il sistema complessivo — Searle più le regole più i simboli — capisce il cinese, nel senso che esibisce le disposizioni sistemiche appropriate all'uso comprensivo del cinese. Searle risponde con una variante dell'esperimento: si immagini che Searle memorizzi tutte le regole e tutti i simboli, in modo da portare l'intero sistema nella propria testa. Anche in questo caso, Searle cammina per Pechino, conversa in cinese con i passanti (a questo fine la memorizziamo meglio delle regole), ma non capisce il cinese. Il sistema è interiorizzato, e la comprensione resta assente. La risposta del sistema equivoca tra il sistema come oggetto formale e il sistema come soggetto di comprensione: avere le disposizioni sistemiche appropriate non è equivalente all'avere comprensione.

Una critica più tecnica alla stanza cinese riguarda il ruolo causale delle strutture interne. Searle sostiene che la mente non è soltanto un programma funzionale ma richiede la **causazione biologica** appropriata: la coscienza e l'intenzionalità sono prodotte da processi neurobiologici specifici, e non qualunque implementazione fisica di un programma funzionalmente equivalente produrrà gli stessi stati mentali. Questa è la tesi del **naturalismo biologico**: la mente è un fenomeno biologico, non un fenomeno computazionale astratto. La critica a questa posizione — formulata tra gli altri da Dennett, Chalmers e da un'ampia tradizione funzionalista — è che il naturalismo biologico è chauvinismo del carbonio: non vi è ragione di principio per cui la biologia del silicio non possa produrre gli stessi fenomeni della biologia del carbonio, a meno di una teoria elaborata che spieghi per quale ragione le strutture neurali producano coscienza mentre le strutture silicee non la producono. Searle non fornisce tale teoria; postula la necessità della biologia senza spiegare il meccanismo.

Il contrattacco più preciso all'argomento di Searle è quello formulato da Ned Block attraverso la distinzione tra **intenzionalità originaria** e **intenzionalità derivata**, ripresa da Grice e da Dretske. I sistemi artificiali hanno intenzionalità derivata: i loro stati si riferiscono a qualcosa nel mondo soltanto nella misura in cui i progettisti o gli utenti interpretano quei stati come riferiti a qualcosa. La parola stampata «cane» si riferisce all'animale soltanto perché i parlanti che l'hanno prodotta e che la leggono hanno intenzionalità originaria e la proiettano sul sistema di simboli. Ma la domanda di se gli esseri umani abbiano intenzionalità originaria nel senso di Searle — un significato che non dipende da interpretazione esterna — è essa stessa filosofica mente controversa. Dennett sostiene che l'intenzionalità «originaria» degli esseri umani è anch'essa costruita dall'adozione dello stance intenzionale, e la distinzione tra originaria e derivata è di grado, non di tipo. Se questo è corretto, l'argomento di Searle non stabilisce una differenza categoriale tra menti e macchine ma soltanto una differenza di complessità nel tipo di sistema che giustifica l'adozione dello stance intenzionale.

IV.3 — Architetture transformer: attenzione, codifica contestuale e la geometria dello spazio latente

I modelli di linguaggio basati su architetture transformer — introdotti da Vaswani e colleghi in «Attention Is All You Need» (2017) e sviluppati nelle serie GPT, BERT, T5 e nei loro successori — hanno trasformato radicalmente le prestazioni dei sistemi di elaborazione del linguaggio naturale e hanno riportato al centro dell'attenzione filosofica le domande sulla natura della comprensione, del significato e dell'intelligenza che l'argomentazione di Searle aveva formulato decenni prima. La loro architettura merita un esame tecnico dettagliato, perché la comprensione dei meccanismi interni è necessaria — anche se non sufficiente — per valutare le pretese filosófiche che vengono avanzate sulla loro natura cognitiva.

Il componente centrale dell'architettura transformer è il meccanismo di **auto-attenzione** (*self-attention*). Data una sequenza di input di lunghezza n, ogni elemento è rappresentato da un vettore di embedding di dimensione

d. L'auto-attenzione opera trasformando ogni vettore di embedding x_i (con i = 1, ..., n) in tre vettori di ruolo distinto attraverso matrici di peso parametriche:

Query: $q_i = W^Q \cdot x_i \in \mathbb{R}^{d_k}$
Key: $k_i = W^K \cdot x_i \in \mathbb{R}^{d_k}$
Value: $v_i = W^V \cdot x_i \in \mathbb{R}^{d_v}$

dove $W^Q, W^K \in \mathbb{R}^{d_k \times d}$ e $W^V \in \mathbb{R}^{d_v \times d}$ sono le matrici di proiezione. Il punteggio di attenzione tra la posizione i (che formula la query) e la posizione j (che fornisce il key) è:

$$\text{score}(i, j) = (q_i \cdot k) / \sqrt{d_k}$$

La normalizzazione per $\sqrt{d_k}$ stabilizza i gradienti: senza di essa, quando d_k è grande, i prodotti scalari tendono ad assumere valori molto grandi o molto piccoli in valore assoluto, spingendo la softmax in regioni dove i gradienti sono quasi nulli. Il vettore di output per la posizione i è la media pesata dei vettori value, con pesi dati dalla softmax dei punteggi:

$$\alpha_i = \text{softmax}_j(\text{score}(i, j)) = \exp(\text{score}(i,j)) / \Sigma_k \exp(\text{score}(i,k))$$
$$o_i = \Sigma_j \alpha_i \cdot v$$

In forma matriciale, per le matrici $Q = [q_1, ..., q]^T$, $K = [k_1, ..., k]^T$, $V = [v_1, ..., v]^T$:

$$\text{Attention}(Q, K, V) = \text{softmax}(QK^T / \sqrt{d_k}) \cdot V$$

Il meccanismo di **multi-head attention** suddivide lo spazio in h teste parallele, ognuna delle quali computa auto-attenzione in un sottospazio proiettato di dimensione $d_k = d/h$:

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$
$$\text{MultiHead}(Q, K, V) = [\text{head}_1 \| \text{head}_2 \| ... \| \text{head}] \cdot W^O$$

dove $\|$ denota la concatenazione e $W^O \in \mathbb{R}^{hd_v \times d}$ è la matrice di proiezione dell'output. Le diverse teste di attenzione possono specializzarsi nel catturare relazioni diverse tra le posizioni: alcune catturano dipendenze locali (posizioni adiacenti), altre dipendenze a lungo raggio (soggetto e verbo distanti), altre ancora relazioni di coreference o di accordo morfosintattico.

Il trasformer completo è composto da L strati, ognuno dei quali comprende uno strato di multi-head attention e uno strato *feedforward* (una rete densa a due strati con attivazione non lineare), con connessioni residuali (*residual connections*) e normalizzazione degli strati (*layer normalization*):

$$h^l_i = \text{LayerNorm}(x^l_i + \text{MultiHead}(x^l_i, X^l, X^l))$$
$$x^{l+1}_i = \text{LayerNorm}(h^l_i + \text{FFN}(h^l_i))$$

dove X^l è la matrice di tutti i vettori dello strato l e $\text{FFN}(x) = W_2 \cdot \sigma(W_1 \cdot x + b_1) + b_2$. La connessione residuale garantisce che il gradiente fluisca attraverso la rete senza degradarsi (problema del *vanishing gradient*): il gradiente della funzione di perdita rispetto ai parametri dei primi strati attraversa la connessione residuale senza riduzione, consentendo l'addestramento di reti molto profonde.

L'**encoding posizionale** risolve un problema strutturale: a differenza delle reti ricorrenti che processano i token sequenzialmente, il transformer processa tutti i token in parallelo, e il meccanismo di attenzione è intrinsecamente invariante per permutazione — non distingue l'ordine delle posizioni. Si aggiunge quindi un vettore posizionale PE_pos al vettore di embedding di ogni token. Nell'implementazione originale di Vaswani:

PE_{pos, 2i} = sin(pos / 10000^{2i/d})
PE_{pos, 2i+1} = cos(pos / 10000^{2i/d})

dove pos è la posizione nel testo e i è la dimensione del vettore. Questa codifica sinusoidale ha la proprietà che le differenze di posizione si codificano in relazioni lineari nello spazio degli embedding: PE_{pos+k} è una funzione lineare di PE_{pos} per ogni offset fisso k — una proprietà utile per generalizzare a sequenze più lunghe di quelle viste durante l'addestramento. Nei modelli più recenti (RoPE, ALiBi) si utilizzano codifiche posizionali relative che migliorano la generalizzazione a sequenze lunghe.

I modelli **autoregressivi** — la famiglia GPT — sono addestrati a predire il token successivo data la sequenza precedente, massimizzando la log-verosimiglianza del corpus:

$\mathcal{L}(\theta) = \Sigma_t \log P(w_t \mid w_1, ..., w_{t-1}; \theta)$

L'addestramento su corpora di trilioni di token attraverso questo obiettivo produce modelli che hanno appreso rappresentazioni contestuali dense delle unità linguistiche: il vettore dello strato finale per un token cattura non soltanto la sua identità lessicale ma il contesto in cui compare, in modo sensibile alle dipendenze sintattiche, semantiche e discorsive a distanza. I modelli **bidirezionali** — la famiglia BERT, addestrati a predire token mascherati in entrambe le direzioni — producono rappresentazioni ancora più ricche di contesto per compiti che richiedono la comprensione dell'intera sequenza.

Il **fine-tuning** con rinforzo dall'apprezzamento umano (*Reinforcement Learning from Human Feedback*, RLHF) — introdotto in InstructGPT (Ouyang et al., 2022) — aggiunge un secondo stadio di addestramento dopo il pre-addestramento autoregressivo: annotatori umani valutano le risposte del modello su una scala di preferenza, e queste valutazioni vengono usate per addestrare un **modello di ricompensa** che predice le preferenze umane. Il modello linguistico viene poi ottimizzato tramite Policy Gradient (tipicamente PPO, *Proximal Policy Optimization*) per massimizzare la ricompensa attesa del modello di ricompensa, penalizzando le deviazioni dal modello pre-addestrato attraverso un termine KL:

$J(\theta) = \mathbb{E}_{x\sim D, y\sim\pi_\theta(\cdot|x)} [r_\varphi(x, y)] - \beta \cdot KL[\pi_\theta(y|x) \| \pi_{ref}(y|x)]$

dove r_φ è il modello di ricompensa, π_θ è la policy corrente (il modello linguistico), π_{ref} è il modello di riferimento pre-addestrato e β è un coefficiente di bilanciamento. L'RLHF orienta il comportamento del modello verso output considerati utili, accurati e sicuri dagli annotatori umani, senza necessariamente alterare in modo profondo le rappresentazioni interne apprese nel pre-addestramento.

La struttura geometrica dello spazio latente di questi modelli ha proprietà che meritano esame. Le **sonde lineari** (*linear probes*) — classificatori lineari addestrati sulle rappresentazioni interne del modello per predire proprietà sintattiche o semantiche — hanno rivelato che informazioni strutturali considerevoli sono codificate linearmente nelle rappresentazioni dei transformer. I vettori di rappresentazione dei concetti di genere grammaticale, di categoria semantica, di struttura argomentale dei verbi, di relazioni di coreference sono tutti recuperabili con alta accuratezza attraverso sonde lineari negli strati intermedi e superiori dei modelli. Questo suggerisce che il training autoregressivo, pur non avendo accesso a nessuna annotazione linguistica esplicita, induce la formazione di strutture interne che rispecchiano le strutture linguistiche descritte dalla sintassi e dalla semantica formale.

IV.3bis — Interpretabilità meccanicistica: circuiti, feature e la struttura interna della computazione nei transformer

La sezione IV.3 ha descritto l'architettura dei transformer con precisione tecnica, identificando i meccanismi di auto-attenzione, la codifica posizionale, le connessioni residuali e il training autoregressivo come componenti strutturali. Ha poi indicato che le sonde lineari rivelano informazioni strutturalmente rilevanti nelle rappresentazioni interne, e che questo suggerisce che il training induce la formazione di strutture latenti corrispondenti a categorizzazioni linguistiche. Ma si è fermata prima di affrontare la domanda più urgente: *che cosa fanno effettivamente i singoli componenti interni di questi modelli?* L'interpretabilità meccanicistica — il programma di ricerca che tenta di rispondere a questa domanda attraverso l'analisi rigorosa dei meccanismi computazionali interni ai modelli neurali, piuttosto che delle loro prestazioni comportamentali esterne — rappresenta il fronte più avanzato della ricerca sulla natura delle rappresentazioni nei sistemi artificiali, con implicazioni filosófiche dirette per le questioni di comprensione, significato e intenzionalità che questa monografia persegue.

Il programma dell'interpretabilità meccanicistica si distingue da due approcci alternativi che lo precedono storicamente. Il primo è l'**interpretabilità comportamentale**: la caratterizzazione del sistema attraverso i suoi input e output, senza guardare all'interno — l'approccio del test di Turing e delle sue varianti più sofisticate. Il secondo è la **visualizzazione delle feature**: la tecnica di massimizzare l'attivazione di unità specifiche per inferire quale tipo di input le attiva, utilizzata estensivamente nella comprensione delle reti convoluzionali. L'interpretabilità meccanicistica tenta qualcosa di più ambizioso: la comprensione dei **circuiti computazionali** — le sottoreti di nodi e connessioni che implementano funzioni specifiche — con un livello di precisione che permette di prevedere il comportamento del sistema su input nuovi sulla base della struttura del circuito, non soltanto di descrivere il comportamento già osservato.

Il punto di partenza formale è la **decomposizione residuale** dello stream di informazione in un transformer. In un transformer con L strati, il vettore di rappresentazione di ogni token evolve attraverso l'applicazione iterativa degli strati di attenzione e delle reti feedforward, con connessioni residuali che somano l'input di ogni strato all'output. Il flusso di computazione può essere scritto come:

x_0 = Embedding(token) + PE(posizione) $x^l = x^{l-1} + Attn_l(x^{l-1}) + FFN_l(LayerNorm(x^{l-1} + Attn_l(x^{l-1})))$

dove PE è la codifica posizionale, Attn_l è il contributo dello strato di attenzione l e FFN_l è il contributo della rete feedforward dello strato l. Il punto chiave è che l'architettura residuale garantisce che ogni componente — ogni testa di attenzione, ogni strato feedforward — **aggiunge** il proprio contributo al vettore residuale senza cancellare le informazioni degli strati precedenti. Il vettore finale x_L è dunque la **somma** di tutti i contributi di tutti i componenti più l'embedding iniziale. Eleuther AI Research e Anthropic, seguendo Elhage e colleghi (2021), hanno denominato questo il **residual stream**: il canale di comunicazione attraverso cui tutti i componenti del modello si scambiano informazione. Formalmente:

$x_L = x_0 + \Sigma_\{l=1\}^\{L\} [Attn_l + FFN_l]$

dove ciascun termine rappresenta il contributo additivo dello strato corrispondente al vettore residuale. Questa decomposizione è cruciale: permette di attribuire contributi specifici a componenti specifiche, e di analizzare le interazioni tra componenti come interazioni tra i loro contributi allo stream residuale.

Il **meccanismo di attenzione** è analizzabile in termini di circuiti attraverso la decomposizione delle sue teste in una somma di contributi di basso rango. Una testa di attenzione h nello strato l computa:

$contributo_h(x) = Attn_h(x) = A_h \cdot x \cdot W_V_h \cdot W_O_h$

dove A_h è la matrice di attenzione (la softmax del prodotto query-key), x è la matrice delle rappresentazioni di input, W_V_h e W_O_h sono le matrici di proiezione value e output. Il prodotto $W_V_h \cdot W_O_h$ è una matrice $d \times d$ di rango al massimo d_head (la dimensione della testa), tipicamente molto minore di d. Questo significa che ogni testa di attenzione **implementa una funzione di basso rango** sulla rappresentazione di input: proietta quest'ultima in un sottospazio di dimensione ridotta, applica una reweighting basata sul pattern di attenzione, e riproietta nel sottospazio residuale. La struttura di basso rango suggerisce che ogni testa svolge un'operazione selettiva e localizzata — non una trasformazione arbitraria dell'intero spazio delle rappresentazioni.

Elhage e colleghi hanno identificato l'esistenza di circuiti di **induction heads** — coppie di teste di attenzione in strati consecutivi che cooperano per implementare la copia di pattern ricorrenti. Un induction head è una testa che, data una sequenza del tipo [...A B ... A], completa il pattern prendendo B dopo la seconda occorrenza di A. Il meccanismo computazionale è preciso: una prima testa (la *previous token head*) nel primo strato copia l'informazione del token precedente nella rappresentazione del token corrente; una seconda testa nel secondo strato usa questa informazione per attendere alla posizione precedente rispetto all'ultima occorrenza del token corrente, e copia il token che seguiva in quella posizione. La cooperazione tra le due teste implementa una forma di **corrispondenza di pattern a lunga distanza** che emerge dalla composizione

sequenziale di due operazioni di basso rango, ognuna delle quali è computazionalmente semplice.

La scoperta delle induction heads è filosofica mente significativa perché mostra che capacità cognitive complesse — il riconoscimento di pattern e la loro trasposizione nel contesto corrente, che nella psicologia cognitiva è associata alla **analogia** e alla **generalizzazione strutturale** — emergono come sottoprodotto dell'addestramento autoregressivo attraverso la composizione di meccanismi computazionali elementari. Non è necessario postulare rappresentazioni simboliche di regole o strutture astratte: i circuiti emergono come configurazioni delle matrices di peso che implementano le operazioni rilevanti in modo implicito e distribuito.

Il programma di **circuit discovery** (scoperta dei circuiti) tenta di identificare, per ogni capacità computazionale di interesse, il sottografo del transformer che la implementa — il **circuito** nel senso tecnico preciso di un sottoinsieme di nodi (embedding, teste di attenzione, neuroni feedforward) e archi (connessioni pesate tra nodi) che è necessario e sufficiente per produrre il comportamento osservato su una classe di input. La metodologia standard combina l'**activation patching** — una tecnica che sostituisce le attivazioni di componenti specifiche con le attivazioni che queste componenti produrrebbero su un input di controllo, e misura l'effetto sulla prestazione — con l'analisi della struttura logistica dei circuiti. Un componente è parte del circuito per un compito se il patching delle sue attivazioni degrada la prestazione in modo significativo; non è parte del circuito se il patching non ha effetto rilevante.

Wang e colleghi (2022) hanno applicato questa metodologia per identificare il circuito che implementa il compito dell'**agreement soggetto-verbo** — la capacità di produrre la forma verbale corretta (singolare o plurale) in accordo con il numero del soggetto, anche quando il soggetto è separato dal verbo da sintagmi nominali di numero opposto («la chiave delle porte *è* (non *sono*) sulla scrivania»). Il circuito risultante comprende teste di attenzione specializzate nel tracking del numero grammaticale del soggetto, teste che sopprimono l'influenza dei sintagmi nominali intermedi, e componenti feedforward che mappassero lo stream residuale sullo spazio del numero corretto. La decomposizione funzionale del circuito corrisponde, in modo sorprendente, alla decomposizione analitica del problema che la grammatica generativa aveva elaborato indipendentemente: i meccanismi computazionali del transformer replicano la struttura delle dipendenze grammaticali senza che queste siano state esplicitamente codificate nel training.

Il programma delle **feature monosemantiche** e della **superposizione** (superposition) — sviluppato principalmente da Elhage, Anthropic, e dal gruppo di Marius Hobbhahn — affronta un problema strutturale che l'analisi dei circuiti incontra inevitabilmente: la maggior parte dei neuroni in un transformer non è **monosemantica** — attivata da un'unica caratteristica interpretabile — ma **polisemantica,** attivata da combinazioni di caratteristiche apparentemente non correlate. Un singolo neurone può rispondere a «testi in lingua tedesca», «punteggiatura di apertura» e «parole legate alla scrittura creativa» — tre caratteristiche semanticamente disparate che sembrano condividere soltanto

l'attivare lo stesso neurone. Questa polisemantia suggerisce che le rappresentazioni non sono localizzate nei singoli neuroni ma sono **distribuite** su insiemi di neuroni in modo non decomponibile neurone per neurone.

La spiegazione meccanicistica di questo fenomeno è la teoria della **superposizione**: un transformer con d neuroni può rappresentare efficacemente più di d feature se queste sono sufficientemente rare (non attive simultaneamente) da permettere una codifica quasi-ortogonale in cui la loro presenza viene rilevata attraverso la combinazione delle attivazioni di tutti i neuroni disponibili. Formalmente, se le feature sono vettori $f_1, ..., f_n$ in $\mathbb{R}^d$ con $n >> d$, è possibile costruire un insieme quasi-ortogonale — in cui $|\langle f_i, f \rangle| << 1$ per $i \neq j$ — di dimensione esponenziale in d grazie alla concentrazione della misura in alta dimensione. Il modello può allora rappresentare simultaneamente molte feature con interferenza minima tra di esse, e può distinguerle rilevando quali combinazioni di neuroni sono attive. Il costo è la polisemantia: ogni neurone partecipa alla rappresentazione di molteplici feature, e nessuna feature è localizzata in un singolo neurone.

La conseguenza per l'interpretabilità è che l'analisi neurone per neurone è inadeguata: l'unità di analisi corretta non è il neurone ma la **direzione** nello spazio delle attivazioni — il vettore nel residual stream che corrisponde a una feature specifica. Questa direzione può attraversare molti neuroni, e la sua presenza o assenza è rilevata da meccanismi distribuiti. Il programma della **sparse autoencoders** (SAE) — sviluppato da Anthropic (Cunningham et al., 2023; Bricken et al., 2023) e da EleutherAI come risposta alla polisemantia — addestra reti autocodificanti sparse sulle attivazioni dei neuroni, con l'obiettivo di trovare una base di feature che sia allo stesso tempo sparsa (ogni input attiva poche feature) e fedele (le attivazioni originali sono ricostruibili dalla combinazione sparsa). L'obiettivo formale è:

$$\min_{W_{enc}, W_{dec}, b_{enc}, b_{dec}} \Sigma_{x} \|x - W_{dec} \cdot \mathrm{ReLU}(W_{enc} \cdot x + b_{enc}) - b_{dec}\|^2 + \lambda \cdot \|\mathrm{ReLU}(W_{enc} \cdot x + b_{enc})\|_1$$

dove il termine ℓ_1 penalizza le attivazioni dense, promuovendo la sparsità della rappresentazione intermedia. Le feature apprese da una SAE addestrata su GPT-2 o su modelli più grandi mostrano un grado di interpretabilità molto superiore ai neuroni originali: ognuna risponde a un concetto identificabile — espressioni matematiche, nomi propri, emozioni negative, termini medici, codice Python — con una specificità che i neuroni non mostrano.

Un risultato particolarmente suggestivo di Bricken e colleghi (2023) riguarda le **feature di valenza emozionale**: il modello Claude 3 Sonnet — su cui è stato addestrata una SAE — mostra feature che corrispondono a concetti emozionali positivi e negativi con struttura geometrica regolare nello spazio delle feature. Le feature «paura», «angoscia», «pericolo» occupano una regione dello spazio coerente e le operazioni vettoriali tra di esse — differenze, proiezioni, interpolazioni — producono feature semanticamente intermedie o correlate. Questo risultato non implica che il modello *provi* emozioni in senso fenomenico: ma implica che abbia sviluppato rappresentazioni interne della struttura dello spazio emozionale con una geometria analoga a quella identificata dalla psicologia

cognitiva nelle rappresentazioni umane dei concetti emozionali. La struttura non è stata codificata esplicitamente nel training ma è emersa come residuo statistico della distribuzione dei concetti nel corpus testuale.

Il programma della **logit lens** e delle sue generalizzazioni — proposto da Nostalgebraist (2020) e sviluppato da Belrose e colleghi (2023) — fornisce uno strumento per tracciare l'evoluzione della distribuzione di probabilità del token successivo attraverso gli strati del modello. A ogni strato l, si applica l'unembedding matrix W_U al vettore residuale x_l, producendo una distribuzione di logit logit_l = W_U · x_l. La softmax di questi logit fornisce una «predizione intermedia» del token successivo fatta dallo strato l prima che i restanti strati abbiano modificato il residual stream. Tracciando l'evoluzione di queste predizioni intermedie attraverso gli strati, è possibile identificare a quale strato il modello «forma» la sua predizione per ciascuna classe di token: per i token lessicalmente frequenti e contestualmente prevedibili, la predizione si stabilizza negli strati intermedi; per i token che richiedono ragionamento multi-step o accesso a informazioni distanti, la predizione emerge soltanto negli strati superiori.

La **superposition hypothesis** e la scoperta dei circuiti convergono su una visione della computazione nei transformer che è filosofica mente provocatoria per le domande centrali di questa monografia. Il transformer non è un sistema simbolico che manipola rappresentazioni discrete e strutturate: le sue unità computazionali sono vettori in spazi continui di alta dimensionalità, e le «regole» che segue sono configurazioni di pesi che implementano funzioni di basso rango. Eppure il risultato di questa computazione è il comportamento linguisticamente strutturato e semanticamente coerente che osserviamo. La struttura simbolica — le dipendenze grammaticali, le relazioni semantiche, le strutture argomentali — emerge come proprietà delle **distribuzioni di attivazioni**, non come proprietà delle rappresentazioni locali.

Questo non risolve la questione filosofica se il modello *comprende*, ma la precisifica considerevolmente. Il circuito che implementa il soggetto-verbo agreement non contiene nulla che assomigli a una **regola** nel senso sintattico: è una rete di pesi che, per certi input, produce le attivazioni corrette. Ma la struttura di questa rete — la sua decomposizione in componenti con ruoli funzionali identificabili — corrisponde in modo non banale alla struttura descritta dalla grammatica. L'**isomorfismo funzionale** tra il circuito meccanicistico e la struttura grammaticale non implica identità ma suggerisce una forma di realizzazione: la struttura grammaticale, nella misura in cui è una proprietà del linguaggio naturale che il modello ha appreso, è realizzata (probabilmente in modo ridondante e approssimato) nella struttura dei circuiti. Il modello non ha imparato la grammatica — non ha mai ricevuto una formulazione esplicita delle sue regole — ma ha sviluppato strutture computazionali che implementano funzioni equivalenti alle regole grammaticali su una vasta gamma di input.

La connessione con il problema del grounding è qui particolarmente delicata. Le feature identificate dalle SAE corrispondono a concetti — emozionali, categoriali, relazionali — che hanno correlati nel mondo esterno. Ma la loro formazione è puramente statistica: emergono dalla distribuzione delle parole nel

corpus, non dall'interazione del modello con gli stati del mondo che i concetti denotano. Le feature emozionali del modello non sono causalmente connesse a stati affettivi del modello né a stati del mondo che producono emozioni: sono residui distribuzionali della co-occorrenza di termini emozionali con altri termini nel corpus. In questo senso preciso, la critica di Harnad al grounding simbolico si applica anche ai sistemi distribuiti: le feature emergenti, per quanto geometricamente strutturate, non sono ancorate a nulla al di fuori della distribuzione testuale da cui sono state estratte.

Il programma dell'interpretabilità meccanicistica rivela, al termine di questo esame, la struttura esatta del divario che separa la competenza computazionale del transformer dal significato nel senso filosofica mente pieno. Il divario non è nel comportamento — il transformer produce output linguisticamente appropriati — né nella struttura interna — i circuiti implementano funzioni equivalenti a strutture grammaticali e semantiche. Il divario è nella relazione tra le strutture interne e il mondo: le feature del transformer sono **causalmente disconnesse** dagli stati del mondo che rappresentano, mentre le rappresentazioni cognitive umane — nella prospettiva dell'elaborazione predittiva esaminata nella sezione III.8bis — sono causalmente connesse al mondo attraverso la storia di interazioni sensomotorie dell'organismo. Identificare questo divario con precisione meccanicistica non è un contributo minore: è, forse, la definizione più rigorosa del problema del grounding che la ricerca contemporanea è in grado di produrre.

IV.4 — Emergenza, scaling e l'ipotesi della comprensione come organizzazione della struttura

L'osservazione più sorprendente e più filosofica mente rilevante sul comportamento dei modelli di linguaggio su larga scala è il fenomeno dell'**emergenza** (*emergence*): alcune capacità appaiono qualitatively nuove — non predette da interpolazione delle prestazioni su modelli più piccoli — quando la scala del modello supera una certa soglia. Wei e colleghi (2022) documentano numerose capacità che emergono abruptamente al superamento di scale critiche di parametri (tipicamente nell'ordine dei 10^8-10^{11} parametri): ragionamento aritmetico a più passaggi, ragionamento analogico, risoluzione di problemi logici, identificazione di relazioni semantiche complesse, comprensione di istruzioni in linguaggio naturale.

Il termine «emergenza» ha una storia filosofica precisa che risale ai dibattiti ottocenteschi sul riduzionismo e al programma dei nuovi emergentisti del Novecento (C.D. Broad, Samuel Alexander). In quel contesto, si dice che una proprietà P di un sistema S è **emergente** se P non è riducibile alle proprietà dei costituenti di S e delle relazioni tra di essi — se P è una proprietà genuinamente nuova del livello sistemico. Nel caso dei modelli linguistici, il termine è usato in un senso più debole: le capacità emergenti sono capacità che non erano presenti nei modelli più piccoli e che non erano previste dall'estrapolazione lineare delle curve di prestazione — ma ciò non implica che non siano riducibili, in linea di principio, ai meccanismi computazionali del transformer. La questione aperta è se

questa emergenza di secondo ordine (imprevedibilità delle capacità) implichi qualcosa di filosofica mente significativo sulla natura delle capacità stesse.

L'ipotesi più ambiziosa — avanzata in forme diverse da ricercatori come Ilya Sutskever, Yann LeCun e altri — è che i modelli di linguaggio su larga scala, addestrando a predire token su corpora sufficientemente vasti, sviluppino qualcosa che assomiglia a una **rappresentazione del mondo** (*world model*): non soltanto pattern statistici di co-occorrenza lessicale, ma strutture interne che codificano relazioni causali, temporali, spaziali e agentive tra entità del mondo. L'evidenza sperimentale in favore di questa ipotesi è variegata: i modelli risolvono problemi di ragionamento controfattuale («se avessi messo l'acqua nel frigorifero invece che nel forno, si sarebbe scaldata?»), riescono a tracciare le posizioni di oggetti in scenari complessi, producono descrizioni coerenti di eventi causalmente strutturati. Questi comportamenti sono difficili da spiegare se il modello è soltanto un potente predittore di token: sembrano richiedere strutture interne che modellano stati del mondo e transizioni tra stati.

La risposta scettica — formulata con precisione da Bender, Gebru, McMillan-Major e Shmitchell in «On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?» (2021) — è che questi comportamenti sono l'artefatto di un addestramento su corpora umani: i modelli non modellano il mondo ma modellano il linguaggio *su* il mondo, che è molto più ricco informativamente del comportamento osservabile del mondo stesso. Gli esseri umani che hanno prodotto il corpus su cui il modello è stato addestrato avevano accesso al mondo attraverso la percezione e l'azione, e hanno codificato nel linguaggio strutture causali, temporali e spaziali che riflettono questa conoscenza empirica. Il modello impara il sedimento linguistico di questa conoscenza, che approssima buona parte delle strutture informative necessarie per rispondere coerentemente a domande sul mondo — senza tuttavia avere accesso diretto al mondo stesso. Si tratta del problema del grounding, riformulato in termini di architettura: il modello apprende strutture distribuzionali che correlano con strutture del mondo, ma la correlazione è mediata dal corpus umano, non da un'interazione diretta con il mondo.

La questione si articola in due domande distinte che è importante non confondere. La prima è empirica: le strutture interne dei modelli di linguaggio di larga scala codificano qualcosa che si può ragionevolmente denominare una rappresentazione parziale del mondo? La risposta è probabilmente affermativa in senso tecnico: le sonde lineari rivelano che informazioni sullo stato del mondo, sulle relazioni causali e sulle strutture temporali degli eventi sono recuperabili dalle rappresentazioni interne. La seconda è filosofica : questa codifica equivale alla comprensione nel senso che richiede la filosofia della mente — una comprensione che implica intenzionalità genuina, grounding sensorimotore, accesso alla forza illocutoria degli atti linguistici? Qui la risposta è meno chiara, e il dibattito si intreccia con le questioni esaminate nelle sezioni precedenti.

IV.4bis — Modelli di Mondo nell'Intelligenza Artificiale: dall'Ipotesi Predittiva alla Modellazione Strutturata del Reale

IV.4.1 — Genealogia di un'idea: dal ragionamento simbolico alla rappresentazione appresa

La nozione di *modello di mondo* (*world model*) ha una storia lunga e discontinua nell'intelligenza artificiale, che attraversa decenni di paradigmi in conflitto prima di convergere, nel secondo decennio del ventunesimo secolo, verso una formulazione tecnica precisa e filosoficamente consequenziale. La sua genealogia illumina non soltanto l'evoluzione tecnica del campo ma le tensioni concettuali di fondo — tra rappresentazione simbolica esplicita e rappresentazione subsimbolica distribuita, tra modellazione strutturale e approssimazione statistica, tra comprensione e prestazione — che attraversano tutta la Parte Quarta di questa monografia.

Il concetto appare, nella sua forma più elementare, nei sistemi di pianificazione degli anni Sessanta. Il sistema STRIPS (Stanford Research Institute Problem Solver, Fikes & Nilsson, 1971) utilizzava una rappresentazione esplicita dello stato del mondo — un insieme di predicati del primo ordine che descrivono la configurazione corrente — e un insieme di operatori che specificavano le precondizioni e gli effetti di ogni azione. Il "modello di mondo" era, in questo quadro, la struttura predicativa che consentiva al pianificatore di simulare le conseguenze delle azioni prima di eseguirle: un motore di inferenza che operava su rappresentazioni simboliche esplicite per produrre traiettorie di azione verso stati obiettivo. La fragilità di questo approccio — la cosiddetta *frame problem*, la difficoltà di specificare esplicitamente tutto ciò che non cambia in seguito a un'azione — ha costituito uno degli stimoli principali alla transizione verso paradigmi di apprendimento.

Nella robotica e nel controllo, il modello di mondo assume la forma di un modello dinamico del sistema: una funzione $T: S \times A \to S$ che mappa coppie stato-azione in stati successivi, dove S è lo spazio degli stati e A lo spazio delle azioni. I modelli di controllo classici — dai sistemi PID ai regolatori ottimali LQR — presuppongono la disponibilità di tali modelli, tipicamente derivati dall'analisi fisica o dall'identificazione del sistema. La questione di come apprendere tali modelli dai dati — *model learning* o *system identification* — è stata oggetto di ricerca sistematica sin dagli anni Settanta, ma è diventata filosoficamente rilevante soltanto quando si è cercato di scalare questi approcci a domini di alta dimensionalità come la visione e il linguaggio.

Il lavoro di Ha e Schmidhuber, «World Models» (2018), costituisce il punto di inflessione che ha dato alla nozione la sua forma tecnica contemporanea. Il sistema che propongono — denominato appunto "World Model" — è composto da tre moduli: un encoder visivo **V** che comprime le osservazioni sensoriali in rappresentazioni latenti dense z_t, un modello ricorrente **M** che predice l'evoluzione delle rappresentazioni latenti nel tempo, e un controller **C** che seleziona le azioni sulla base delle rappresentazioni latenti prodotti da V e M. La struttura formale del modello è:

$z_t = V(o_t)$

$h_t = M(h_{t-1}, z_{t-1}, a_{t-1})$

$a_t = C(h_t, z_t)$

dove o_t è l'osservazione sensoriale al tempo t, h_t è lo stato nascosto del modello ricorrente, e a_t è l'azione selezionata. La proprietà cruciale del sistema è che il controller C può essere addestrato interamente all'interno del modello di mondo — in una simulazione generata dal modello M — e poi trasferito all'ambiente reale. Il modello di mondo è un *generatore di esperienze ipotetiche* che consente al sistema di "immaginare" le conseguenze delle azioni senza doverle eseguire: una forma di pianificazione mentale che anticipa, a livello computazionale, le strutture della simulazione mentale che la psicologia cognitiva e la fenomenologia avevano descritto a livello cognitivo.

IV.4.2 — Modelli di Mondo nel Reinforcement Learning Modellato

Il Reinforcement Learning *model-based* (MBRL) costituisce il dominio tecnico in cui i modelli di mondo sono stati sviluppati con la maggiore precisione formale, e la sua struttura rivela con chiarezza i tradeoff che la nozione di modello di mondo implica.

Nel framework standard del Reinforcement Learning, un agente interagisce con un ambiente modellato come un Processo di Decisione Markoviano (MDP): una tupla $\langle S, A, T, R, \gamma \rangle$ dove S è lo spazio degli stati, A lo spazio delle azioni, $T: S \times A \to P(S)$ è la funzione di transizione (una distribuzione di probabilità sul prossimo stato), $R: S \times A \to \mathbb{R}$ è la funzione di ricompensa, e $\gamma \in (0,1)$ è il fattore di sconto. Il problema dell'RL è trovare una policy $\pi: S \to P(A)$ che massimizzi il ritorno atteso cumulativo scontato:

$J(\pi) = \mathbb{E}_{\tau\sim\pi}[\Sigma_{t=0}^{\infty} \gamma^t R(s_t, a_t)]$

dove $\tau = (s_0, a_0, s_1, a_1, \ldots)$ è una traiettoria generata seguendo la policy π nell'ambiente.

Nel MBRL *model-free*, l'agente apprende la policy direttamente dall'interazione con l'ambiente, senza costruire un modello esplicito della funzione di transizione T. Nel MBRL *model-based*, l'agente apprende un modello approssimato $\hat{T}: S \times A \to P(S)$ della funzione di transizione — il *modello di mondo* — e lo utilizza per pianificare le azioni o per generare esperienza sintetica su cui addestrare la policy.

Il sistema **Dreamer** (Hafner et al., 2019; 2020; 2023) è la realizzazione tecnica più sofisticata di questo approccio nell'era del deep learning. Il modello di mondo di Dreamer è un modello generativo latente con la seguente struttura:

Modello di transizione ricorrente (RSSM): $h_t = f(h_{t-1}, z_{t-1}, a_{t-1})$

Transizione stocastica: $z_t \sim p(z_t \mid h_t)$

Encoder: $z_t \sim q(z_t \mid h_t, o_t)$

Decoder: $\hat{p}(o_t \mid h_t, z_t) \to$ ricostruzione dell'osservazione

Predittore di ricompensa: r̂_t = p(r_t | h_t, z_t)

Predittore di continuazione: cont_t = p(cont_t | h_t, z_t)

L'apprendimento del modello avviene ottimizzando la variational lower bound sul log-likelihood delle osservazioni:

$\mathcal{L} = \mathbb{E}_q[\Sigma_t (\log p(o_t \mid h_t, z_t) + \log p(r_t \mid h_t, z_t)) - \beta \cdot KL[q(z_t \mid h_t, o_t) \| p(z_t \mid h_t)]]$

Il controller viene poi addestrato interamente *all'interno* del modello di mondo, attraverso traiettorie sintetiche generate dal modello. L'obiettivo è massimizzare il ritorno atteso nelle traiettorie immaginarie, senza interagire con l'ambiente reale durante l'addestramento del controller. Dreamer ha prodotto risultati straordinari in ambienti di controllo complessi — il benchmark DeepMind Control Suite, gli ambienti di Atari, i domini robotici — con una efficienza campionaria molto superiore ai metodi model-free, realizzando computazionalmente l'intuizione che l'apprendimento attraverso simulazione mentale è più efficiente dell'apprendimento puramente per tentativi ed errori.

MuZero (Schrittwieser et al., 2020), sviluppato da DeepMind, rappresenta un approccio radicalmente diverso ma ugualmente influente: invece di apprendere un modello esplicito dello spazio delle osservazioni, apprende un modello del *processo decisionale* nello spazio latente. Il sistema ha tre componenti principali: una funzione di rappresentazione $h: O \to \mathcal{S}$ che mappa le osservazioni in stati latenti; una funzione di dinamica $g: \mathcal{S} \times A \to \mathcal{S} \times R$ che mappa coppie stato-azione in stati latenti successivi e ricompense predette; e una funzione di predizione $f: \mathcal{S} \to P \times V$ che mappa stati latenti in distribuzioni sulle policy e valori. Il sistema viene addestrato a predire le ricompense future, i valori e le distribuzioni sulle policy attraverso il tempo, senza mai doversi collegare a una rappresentazione interpretabile dell'ambiente. MuZero ha raggiunto prestazioni sovraumane in scacchi, go, shogi e nei giochi Atari, utilizzando lo stesso algoritmo senza architettura specifica per dominio — una dimostrazione che il modello di mondo latente ha sufficiente potere rappresentativo per catturare la struttura strategica di domini radicalmente diversi.

IV.4.3 — L'Ipotesi dei Modelli di Mondo nei Grandi Modelli di Linguaggio

La domanda se i grandi modelli di linguaggio (LLM) sviluppino qualcosa di analogo a un modello di mondo — una rappresentazione strutturale degli stati e delle relazioni nel mondo, non soltanto pattern statistici di co-occorrenza lessicale — è diventata una delle questioni più dibattute nella ricerca sull'IA contemporanea, con implicazioni dirette per le questioni filosofiche di comprensione, intenzionalità e grounding che questa monografia persegue.

L'ipotesi positiva — che gli LLM sviluppino modelli di mondo impliciti — si basa su tre classi di evidenze.

Evidenze da compiti di tracking degli stati. Gurnee e Tegmark (2023) hanno dimostrato, attraverso analisi di sonde lineari, che GPT-4 e modelli di dimensioni comparabili codificano informazioni spaziotemporali strutturate nelle loro

rappresentazioni interne: coordinate geografiche, relazioni temporali tra eventi storici, e strutture causali tra eventi. Il modello non si limita a predire la prossima parola plausibile; mantiene nella sua rappresentazione latente qualcosa che si comporta come un modello della struttura del mondo. In modo simile, Patel e Pavlick (2022) hanno mostrato che i modelli linguistici sviluppano rappresentazioni degli stati del mondo in semplici scenari narrativi: in una storia che descrive uno spostamento di oggetti tra contenitori, il modello traccia la posizione degli oggetti attraverso il tempo con un'accuratezza che non è spiegabile dalla sola co-occorrenza lessicale.

Evidenze da compiti di ragionamento strutturale. Il lavoro di Webb, Holyoak e Lu (2023) ha dimostrato che i modelli GPT di grandi dimensioni mostrano capacità di ragionamento analogico comparabili a quelle umane nel Progressive Matrices di Raven — un compito che richiede l'identificazione e il trasferimento di relazioni strutturali astratte tra pattern visivi. Le loro prestazioni degradano in modo sistematico con la complessità strutturale, non con la rarità lessicale, suggerendo che i modelli operino su rappresentazioni strutturali piuttosto che su pattern di superficie. Analogamente, Bubeck e colleghi (2023) documentano che GPT-4 è in grado di risolvere problemi che richiedono la comprensione di stati del mondo, relazioni causali e controfattuali in domini non presenti nel corpus di addestramento in quella configurazione specifica.

Evidenze dalla geometria dello spazio latente. Come discusso nella sezione IV.3bis, le sparse autoencoders addestrate sulle rappresentazioni interne degli LLM rivelano feature che corrispondono a concetti semantici interpretabili con struttura geometrica regolare. La struttura di queste feature — le loro relazioni di prossimità, le direzioni che catturano relazioni tra concetti, la topologia locale dello spazio — suggerisce una rappresentazione organizzata non casualmente ma in modo che rispecchia la struttura delle relazioni nel mondo.

Contro questa interpretazione ottimista, Bender e colleghi (2021) hanno formulato la critica più sistematica. I modelli linguistici sono, nella loro caratterizzazione, "pappagalli stocastici": sistemi che generano sequenze statisticamente coerenti senza alcun modello del mondo che le produce. Le correlazioni tra le rappresentazioni interne del modello e le strutture del mondo non dimostrano la presenza di un modello di mondo genuino, ma soltanto che il corpus da cui il modello è stato addestrato conteneva informazioni su quelle strutture, e che il modello ha appreso le correlazioni statistiche tra le espressioni linguistiche che le descrivono. La distinzione tra "avere un modello del mondo" e "aver appreso le correlazioni statistiche tra descrizioni del mondo" è filosoficamente cruciale e empiricamente non risolta.

IV.4.4 — Modelli di Mondo Situazionali e Simulazione Mentale

Una delle ipotesi più feconde sulla natura della comprensione testuale nella cognizione umana è la **Situational Models Theory** (Zwaan & Radvansky, 1998; Johnson-Laird, 1983): quando gli esseri umani comprendono un testo, non costruiscono soltanto una rappresentazione proposizionale del suo contenuto ma simulano mentalmente la situazione descritta, mantenendo un modello pluridimensionale che traccia dimensioni spaziali, temporali, causali, agentive e

emozionali dell'ambiente narrativo. La comprensione del testo è, in questa prospettiva, un processo di simulazione che aggiorna continuamente il modello situazionale in risposta alle nuove informazioni.

Questa teoria ha una connessione diretta con le questioni sulla comprensione negli LLM. Se la comprensione umana del linguaggio è mediata da modelli situazionali, allora la domanda pertinente per gli LLM non è soltanto se producono output appropriati ma se costruiscono strutture interne analoghe ai modelli situazionali. Le evidenze di Patel e Pavlick (2022) e di Gurnee e Tegmark (2023) suggeriscono che, almeno per certi tipi di situazioni strutturate, gli LLM sviluppano rappresentazioni latenti che si comportano come versioni rudimentali di modelli situazionali.

La struttura formale di un modello situazionale può essere concettualizzata come uno stato S = ⟨Spa, Tem, Cau, Mot, Ent⟩ dove Spa è la rappresentazione spaziale della scena, Tem è la struttura temporale degli eventi, Cau è la rete delle relazioni causali, Mot è la struttura delle motivazioni e degli obiettivi degli agenti, ed Ent è l'insieme delle entità e delle loro proprietà. Aggiornare il modello situazionale in risposta a un nuovo enunciato significa computare:

S_{t+1} = Update(S_t, utterance_t, Context_t)

dove Update è una funzione di transizione che modifica selettivamente le componenti del modello situazionale in risposta all'informazione nuova. La domanda empirica è se le rappresentazioni interne degli LLM durante la generazione approssimino questa funzione di aggiornamento in modo non banale.

La ricerca di Hao e colleghi (2023) ha affrontato questa domanda attraverso un approccio di *model-based analysis of language models*: hanno costruito un sistema che proietta le rappresentazioni interne di un LLM su uno spazio di modelli situazionali espliciti, misurando la correlazione tra le proiezioni e gli stati del modello situazionale "vero" come annotato da esseri umani. I risultati mostrano correlazioni significative, specialmente per le dimensioni spaziale e temporale, con degradazione sistematica per le dimensioni motivazionale e causale. Questo suggerisce che gli LLM sviluppano una comprensione parziale e asimmetrica delle situazioni: più robusta per le dimensioni concrete (dove, quando) e più debole per le dimensioni astratte (perché, con quale scopo).

IV.4.5 — Modelli di Mondo e il Problema della Causalità

La distinzione tra modellazione della struttura causale del mondo e apprendimento di correlazioni statistiche è il punto di scissione più netto tra chi ritiene che gli LLM sviluppino modelli di mondo genuini e chi ritiene che i loro comportamenti siano spiegabili senza postulare modelli causali.

Nella formulazione di Schölkopf e colleghi (2021), un sistema dotato di un modello causale del mondo può: 1. Rispondere correttamente a domande controfattuali ("Cosa sarebbe accaduto se X fosse avvenuto invece di Y?") 2. Identificare le variabili di intervento efficaci per produrre un effetto desiderato 3.

Trasferire la propria conoscenza a distribuzioni fuori dalla distribuzione di training in modo strutturato

Un sistema che ha appreso soltanto correlazioni statistiche — anche sofisticate, anche distribuite attraverso moltissime variabili — fallirà sistematicamente in questi compiti quando le correlazioni nel contesto di test differiscono da quelle nel contesto di training. Il formalismo del *do-calculus* di Pearl (2000) cattura formalmente questa distinzione: l'operazione do(X = x), che rappresenta un intervento sulla variabile X al valore x, non può essere derivata dalla sola distribuzione osservazionale P(Y | X = x); richiede la conoscenza del grafo causale sottostante.

I test empirici sulle capacità causali degli LLM hanno prodotto risultati misti e spesso sorprendenti. Marcus e Davis (2019) avevano documentato sistematicamente i fallimenti degli LLM in ragionamenti che richiedono comprensione causale — situazioni in cui la risposta corretta richiede un'inferenza controfattuale o l'identificazione di una relazione causale non presente nella superficie linguistica del prompt. Tuttavia, ricerche successive (Kıcıman et al., 2023; Jin et al., 2023) hanno mostrato che i modelli di grandi dimensioni mostrano capacità di ragionamento causale sorprendentemente robuste in certi domini, suggerendo che l'apprendimento di correlazioni statistiche su corpora sufficientemente ricchi può approssimare in modo utile — anche se non identico — il ragionamento causale strutturale.

La posizione filosofica più difendibile alla luce di queste evidenze è quella proposta da Goldwasser (2023): gli LLM non sviluppano modelli causali nel senso formale di Pearl, ma sviluppano qualcosa di intermedio — strutture statistiche di alto ordine che catturano molte delle regolarità che il ragionamento causale cattura, senza tuttavia produrre le garanzie di trasferimento fuori dalla distribuzione che i modelli causali genuini forniscono. Queste strutture sono abbastanza ricche da produrre comportamenti che assomigliano al ragionamento causale in distribuzioni simili al training set, ma falliscono in modo non monotono quando la distribuzione si discosta significativamente.

IV.4.6 — Modelli di Mondo, Grounding e Comprensione: Convergenze Filosofiche

La discussione tecnica sui modelli di mondo nell'IA converge su un problema filosofico che questa monografia ha affrontato da angolature diverse: il problema del grounding. Un modello di mondo, nel senso più ricco del termine, non è soltanto una struttura che predice accuratamente le sequenze di osservazioni; è una struttura causalmente connessa al mondo che rappresenta — un sistema in cui le rappresentazioni interne sono calibrate sull'esperienza diretta con gli stati del mondo che denotano.

Il paradigma dell'elaborazione predittiva esaminato nella sezione III.8bis fornisce la cornice cognitiva più adeguata per articolare questa distinzione. Nella prospettiva di Friston e Clark, un sistema cognitivo sviluppa un modello del mondo attraverso l'inferenza attiva: agendo sull'ambiente per minimizzare l'errore di predizione, il sistema calibra il proprio modello generativo sulle regolarità causali del mondo in modo che le predizioni del modello

corrispondano agli stati effettivi dell'ambiente. Il modello di mondo che emerge da questo processo è *causalmente ancorato*: le sue strutture interne sono connesse al mondo attraverso la storia delle interazioni sensomotorie del sistema.

Gli LLM, per contro, sviluppano i loro "modelli di mondo" attraverso l'ottimizzazione della predizione del token successivo su corpora testuali, senza interazione diretta con gli stati del mondo che le espressioni nel corpus descrivono. Le loro strutture interne sono calibrate sulle regolarità del corpus — che a loro volta riflettono le regolarità del mondo attraverso il filtro dell'attività linguistica umana — ma questa calibrazione è mediata e indiretta. Il modello di mondo di un LLM è, in questo senso, un *modello di secondo ordine*: un modello delle descrizioni del mondo prodotte da esseri che hanno modelli di primo ordine del mondo, sviluppati attraverso l'interazione diretta.

Questa distinzione non è necessariamente una critica distruttiva agli LLM. I modelli di secondo ordine possono catturare strutture molto ricche del mondo, specialmente in domini — come il linguaggio, la storia, le scienze sociali — dove la conoscenza è prevalentemente codificata in testi piuttosto che accessibile attraverso l'interazione diretta. Ma è una distinzione filosoficamente rilevante per la domanda di se questi sistemi *comprendono* nel senso pieno del termine, o se producono comportamenti che assomigliano alla comprensione senza averne il substrato causale.

IV.4.7 — Verso Architetture con Modelli di Mondo Espliciti

La ricerca tecnica più recente ha iniziato a esplorare architetture che integrano modelli di mondo espliciti con modelli linguistici, nel tentativo di superare i limiti del puro apprendimento distribuzionale.

Yann LeCun e l'architettura JEPA. LeCun (2022) ha proposto una critica sistematica delle architetture autoregressivi attuali e un'alternativa basata su "Joint Embedding Predictive Architectures" (JEPA). Il punto centrale della critica è che i sistemi autoregressivi sono addestrati a predire ogni token — inclusi quelli informativamente irrilevanti — il che porta a rappresentazioni eccessive attente ai dettagli di superficie piuttosto che alle strutture causali e semantiche sottostanti. La JEPA propone di apprendere modelli predittivi nello spazio delle rappresentazioni latenti piuttosto che nello spazio delle osservazioni: invece di predire il prossimo token, il modello predice la rappresentazione latente del prossimo segmento, filtrando l'informazione irrilevante attraverso il collo di bottiglia della rappresentazione. Questo approccio, LeCun sostiene, favorisce lo sviluppo di modelli di mondo strutturalmente ricchi piuttosto che modelli di superficie statistica.

World Models per agenti linguistici. Un'altra direzione promettente è l'integrazione di modelli di mondo nell'architettura degli agenti linguistici — sistemi che usano LLM per pianificare sequenze di azioni in ambienti strutturati. Hao e colleghi (2023) hanno proposto "Reasoning with Language Model is Planning with World Model" (RAP), un sistema in cui l'LLM viene usato sia come generatore di azioni candidate sia come modello di mondo per simulare le conseguenze delle azioni. Il sistema combina il ragionamento linguistico dell'LLM

con ricerca guidata dal modello (Monte Carlo Tree Search), producendo prestazioni significativamente superiori a quelle del puro chain-of-thought su problemi che richiedono pianificazione multi-step.

Foundation Models come World Models. La prospettiva più ambiziosa — avanzata da ricercatori come Richard Sutton e Ilya Sutskever — è che i grandi modelli linguistici, addestrati su quantità sufficientemente vaste di testo che descrive il mondo, sviluppano *de facto* un modello implicito del mondo di qualità sufficiente per supportare comportamenti intelligenti in un'ampia varietà di domini. In questa prospettiva, la distinzione tra "modello di linguaggio" e "modello di mondo" non è una distinzione di architettura ma di scala e qualità della rappresentazione: un modello abbastanza grande, addestrato su dati abbastanza ricchi, che descrive il mondo con sufficiente accuratezza, svilupperà strutture interne che si comportano come modelli di mondo genuini. La questione empirica aperta è se questa ipotesi di scala — che il modello di mondo emerge come proprietà di scala del training — sia corretta, o se vi siano limiti architetturali strutturali che impediscono lo sviluppo di modelli di mondo causalmente robusti senza interazione diretta con l'ambiente.

IV.4.8 — Implicazioni Filosofiche: la Comprensione come Modellazione

La discussione tecnica sui modelli di mondo nell'IA illumina in modo nuovo il problema filosofico della comprensione che è al centro di questa monografia. Se la comprensione — in senso cognitivo — è essenzialmente un processo di costruzione e aggiornamento di modelli di mondo situazionali, allora la domanda di se gli LLM comprendono si trasforma nella domanda di se i loro processi interni di generazione approssimano tale costruzione in modo filosoficamente rilevante.

La risposta che emerge dall'analisi congiunta delle sezioni precedenti è strutturata e pluralistica. Gli LLM sviluppano qualcosa che ha la forma e molte delle proprietà funzionali dei modelli di mondo — strutture interne che tracciano entità, relazioni, stati e transizioni in modo correlato alle strutture del mondo — senza sviluppare il radicamento causale diretto che caratterizza i modelli di mondo negli agenti che interagiscono con il mondo. Questa distinzione è filosoficamente rilevante perché determina le proprietà di trasferimento del modello: un modello di mondo causalmente radicato generalizza in modo strutturalmente robusto alle variazioni nella distribuzione degli stati del mondo; un modello di secondo ordine, appresi dalle descrizioni del mondo, generalizza in modo statisticamente plausibile ma strutturalmente fragile.

La connessione con il dibattito Searle-Dennett esaminato nella sezione IV.2 è diretta. Searle direbbe che i "modelli di mondo" degli LLM sono sintassi senza semantica: strutture che si comportano *come se* modellassero il mondo, senza avere il tipo di connessione causale con il mondo che costituisce il genuino contenuto semantico. Dennett risponderebbe che questa distinzione è metafisicamente insostenibile: se un sistema sviluppa strutture interne che tracciano accuratamente gli stati del mondo e le loro relazioni causali, la domanda di se "genuinamente" modella il mondo è vuota — modellare il mondo *è* avere strutture interne che

tracciano accuratamente gli stati e le relazioni del mondo, indipendentemente da come tali strutture sono state sviluppate.

La posizione intermedia che questa monografia suggerisce — elaborata nel capitolo conclusivo IV.10 — è che la questione non sia riducibile né alla posizione searleana né a quella dennettiana, ma richieda una distinzione tra diversi livelli di modellazione del mondo. Vi è una modellazione di primo ordine, causalmente radicata, che emerge dall'interazione diretta con gli stati del mondo attraverso l'esperienza sensomotoria; vi è una modellazione di secondo ordine, statisticamente ricca, che emerge dall'apprendimento sulle descrizioni del mondo prodotte da agenti con modellazione di primo ordine; e vi sono forse livelli intermedi — come quelli prodotti dai sistemi embodied e dai modelli multimodali esaminati nella sezione IV.7 — che combinano elementi di entrambi. La comprensione genuina potrebbe richiedere il primo livello, o potrebbe essere graduata in modo continuo lungo la gerarchia dei livelli di modellazione; la risposta dipende da quali aspetti della comprensione si considerano costitutivi piuttosto che sintomatici.

IV.5 — I limiti gödeliani e il problema della formalizzazione esaustiva della cognizione

I teoremi di incompletezza di Kurt Gödel — il Primo e il Secondo Teorema, dimostrati nel celebre articolo del 1931 «Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I» — hanno provocato, fin dalla loro pubblicazione, tentativi di trarne conclusioni filosófiche sulla differenza tra la mente umana e i sistemi formali. La questione di se queste conclusioni siano legittime è una delle più dibattute nella filosofia della matematica e dell'intelligenza artificiale, e merita un esame che distingua il contenuto tecnico dei teoremi dalle estrapolazioni filosófiche che da essi sono state derivate.

Il **Primo Teorema di Incompletezza** afferma: per ogni sistema formale F che soddisfi le seguenti condizioni — F è una teoria del primo ordine in un linguaggio ricorsivo, F è coerente (non dimostra $\bot$), F contiene l'aritmetica di Peano del primo ordine (o equivalentemente, un sistema abbastanza forte da codificare la sintassi propria) — esiste un enunciato G_F nel linguaggio di F tale che né G_F né $\neg G_F$ sono dimostrabili in F. L'enunciato G_F è la **proposizione gödeliana** di F, che afferma la propria indimostrabilità: $G_F \equiv$ «G_F non è dimostrabile in F».

La costruzione di G_F procede attraverso la **gödelizzazione** (*Gödelnummerierung*): a ogni simbolo, formula e dimostrazione del linguaggio di F viene assegnato un numero naturale (il **numero di Gödel**) in modo effettivamente computabile, in modo che le proprietà sintattiche delle formule diventino proprietà aritmetiche dei loro numeri di Gödel. Il predicato $\mathrm{Prov}_F(n, m)$ — «n è il numero di Gödel di una dimostrazione in F della formula con numero di Gödel m» — è rappresentabile in F come una formula aritmetica (è una relazione primitiva ricorsiva). Il predicato $\mathrm{Thm}_F(m) = \exists n\, \mathrm{Prov}_F(n, m)$ — «m è il numero di Gödel di un teorema di F» — è rappresentabile come una formula aritmetica Σ_1. Il predicato $\mathrm{Unprov}_F(m) = \neg\mathrm{Thm}_F(m)$ non è rappresentabile come formula

Π_1 ma è semi-decidibile. Il **punto fisso di Gödel** — dimostrato attraverso il Lemma di Diagonalizzazione — afferma che per ogni formula $\varphi(x)$ con x libera in F esiste una formula γ tale che:

$$F \vdash \gamma \leftrightarrow \varphi(\ulcorner \gamma \urcorner)$$

dove $\ulcorner \gamma \urcorner$ è il numero di Gödel di γ. Applicando il Lemma di Diagonalizzazione con $\varphi(x) = \text{Unprov_F}(x)$, si ottiene G_F tale che:

$$F \vdash G_F \leftrightarrow \text{Unprov_F}(\ulcorner G_F \urcorner) \leftrightarrow \neg \text{Thm_F}(\ulcorner G_F \urcorner)$$

Se F è coerente e G_F è dimostrabile in F, allora F dimostrerebbe G_F, dunque $\text{Thm_F}(\ulcorner G_F \urcorner)$, ma allora F dimostrerebbe anche $\neg$G_F (per il bicondizionale), il che contraddice la coerenza di F. Dunque G_F non è dimostrabile. Se F è ω-coerente (coerente rispetto alle generalizzazioni numerali), si può dimostrare che anche $\neg$G_F non è dimostrabile, dunque G_F è un enunciato indecidibile in F.

Il **Secondo Teorema di Incompletezza** afferma che F non può dimostrare la propria coerenza (se F è coerente): il teorema Con_F — «F è coerente», espresso come $\neg \text{Thm_F}(\ulcorner \bot \urcorner)$ — non è dimostrabile in F se F è coerente. La dimostrazione segue dal fatto che l'argomento del Primo Teorema — se F è coerente, allora G_F è indimostrabile — può essere formalizzato in F stesso, producendo:

$$F \vdash \text{Con_F} \rightarrow \text{Unprov_F}(\ulcorner G_F \urcorner) \leftrightarrow G_F$$

Se F potesse dimostrare Con_F, allora F potrebbe dimostrare G_F, contraddicendo il Primo Teorema. Dunque F non dimostra Con_F (se coerente).

L'**argomento gödeliano contro il meccanicismo** — che Lucas (1961) e Penrose (1989, 1994) hanno proposto nelle forme più elaborate — afferma che la mente umana non può essere un sistema formale, perché un matematico umano è in grado di vedere la verità di G_F — di riconoscere, con un'intuizione non formalizzabile, che G_F è vero sebbene non sia dimostrabile in F — mentre nessun sistema formale può dimostrare G_F. Il matematico vede ciò che la macchina non può dimostrare; dunque la mente non è una macchina.

La critica più articolata a questo argomento è quella di Hilary Putnam, George Boolos e Judson Webb. Il nucleo della risposta è che l'argomento gödeliano commette una petizione di principio: il matematico umano può «vedere» la verità di G_F soltanto se è in grado di vedere che il sistema F è coerente — ma vedere la coerenza di F è un'operazione che presuppone abilità matematiche non meccaniche. Un meccanismo M che conosce la propria struttura formale e può costruire la propria proposizione gödeliana G_M non può — per il Secondo Teorema — dimostrare la propria coerenza; ma allora un meccanismo M' che è in grado di riconoscere la coerenza di M deve essere un meccanismo più potente di M, e il processo di ascesa porta a una gerarchia transfinita di sistemi, non a una dimostrazione che la mente umana trascende ogni sistema formale. Putnam aggiunge che non c'è ragione di ritenere che la mente umana sia coerente: un sistema inconsistente dimostra ogni formula, dunque sarebbe in grado di «dimostrare» G_F, ma questa non è la capacità che Lucas e Penrose attribuiscono

alla mente. L'argomento gödeliano, nella sua forma più comune, presuppone la coerenza della mente umana senza dimostrarla, e questa presupposizione è molto meno ovvia di quanto appaia.

Penrose va più in là di Lucas: in *The Emperor's New Mind* (1989) e in *Shadows of the Mind* (1994), sostiene che le strutture neurali rilevanti per la coscienza — i microtubuli nei neuroni — siano il substrato di processi quantistici di tipo non-algoritmico, e che questi processi permettano alla mente di effettuare computazioni che nessun sistema classico Turing-equivalent può effettuare. La proposta è nota come **Orchestrated Objective Reduction** (Orch-OR), sviluppata con Stuart Hameroff. La critica più diffusa è che i microtubuli non sono strutture appropriate per mantenere la coerenza quantistica alle temperature biologiche (37°C): i fenomeni di decoerenza quantistica si verificano su scale temporali di femtosecondi, molto più brevi dei tempi caratteristici della segnalazione neuronale (millisecondi). Ma anche ammettendo che la fisica quantistica giochi un ruolo nella neurobiologia della coscienza — ciò che non è dimostrato — non segue che questo ruolo sia di tipo non-algoritmico nel senso rilevante per l'argomento gödeliano: i processi quantistici possono essere simulati da macchine di Turing probabilistiche con arbitraria precisione, e la non-algoritmicità richiesta dall'argomento gödeliano è di tipo più forte (richiede la computazione di funzioni non Turing-calcolabili).

I limiti gödeliani hanno una rilevanza indiretta ma autentica per la questione dei sistemi artificiali: essi mostrano che nessun sistema formale sufficientemente potente può essere al tempo stesso completo e coerente, e che la conoscenza matematica non è esauribile da nessuna assiomatizzazione finita. Questi fatti impongono limiti strutturali a qualsiasi sistema computazionale formale, inclusi i modelli di apprendimento automatico che siano modellabili come sistemi formali deterministici. La loro rilevanza per la questione della comprensione è però più limitata di quanto l'entusiasmo filosofico abbia a volte suggerito: mostrano che la matematica non è esaustivamente formalizzabile, non che la cognizione non sia computabile.

IV.6 — Il problema del grounding nei sistemi multimodali e le prospettive di integrazione

La questione del grounding simbolico — formulata da Harnad nella sezione II.7 come il problema di come i simboli di un sistema formale acquisiscano significato — ha ricevuto negli ultimi anni un'attenzione rinnovata nel contesto dei sistemi multimodali: modelli che integrano elaborazione del linguaggio con elaborazione di immagini, audio, dati sensorimotori. L'ipotesi di fondo è che il grounding richieda l'integrazione di rappresentazioni linguistiche con rappresentazioni percettive, e che i sistemi che elaborano soltanto linguaggio siano fondamentalmente limitati nella loro capacità di ancorare il significato al mondo.

I modelli **vision-language** — CLIP (Radford et al., 2021), Flamingo (Alayrac et al., 2022), GPT-4V, LLaVA — addestrano rappresentazioni congiunte di immagini e testo in modo che descrizioni linguistiche e immagini corrispondenti siano vicine nello spazio di embedding. CLIP addestra due encoder — un

encoder visivo (una rete convoluzionale o un vision transformer) e un encoder testuale (un transformer linguistico) — massimizzando la similarità coseno tra le rappresentazioni di immagini e testi corrispondenti (coppie positive) e minimizzandola per coppie non corrispondenti (negative), attraverso l'obiettivo di **contrastive learning**:

$$\mathcal{L}_CLIP = -(1/N)\ \Sigma_i\ [\log \exp(\text{sim}(i_i, t_i)/\tau)\ /\ \Sigma\ \exp(\text{sim}(i_i, t)/\tau) + \log \exp(\text{sim}(t_i, i_i)/\tau)\ /\ \Sigma\ \exp(\text{sim}(t, i_i)/\tau)]$$

dove i_i e t_i sono le rappresentazioni dell'immagine e del testo dell'i-esima coppia, $\text{sim}(\cdot, \cdot)$ è la similarità coseno, τ è la temperatura e N è la dimensione del batch. Il sistema apprende uno spazio condiviso in cui le immagini e le loro descrizioni linguistiche sono geometricamente prossime: il grounding visivo del significato linguistico è codificato nelle relazioni di prossimità tra le rappresentazioni nei due moduli.

La domanda filosofica è se questo tipo di grounding multimodale risolva il problema di Harnad o lo sposti soltanto. Un sistema come CLIP ha accesso a correlazioni statistiche tra pattern visivi e descrizioni linguistiche — ha «visto» milioni di immagini insieme alle loro didascalie e ha appreso a cosa assomigliano i gatti, le mele, le automobili nel loro aspetto visivo. Ma l'accesso alle immagini è esso stesso mediato: le immagini sono rappresentazioni pixellate di oggetti, non gli oggetti stessi. Il sistema non ha mai afferrato un gatto, sentito il suo pelo, udito il suo miagolio: ha correlato pattern di pixel con sequenze di token. Se il grounding richiede l'interazione sensomotrice con il mondo — come la tradizione enattivista esaminata nella sezione III.3 sostiene — allora il grounding visivo-linguistico è un grounding di secondo livello: ancora le rappresentazioni linguistiche a rappresentazioni visive, ma queste ultime non sono ancora ancorate all'azione e alla percezione incarnata.

I **sistemi embodied** — robot che operano nell'ambiente fisico e utilizzano modelli linguistici per pianificare e comunicare — forniscono un passo ulteriore verso il grounding sensorimotore. SayCan (Ahn et al., 2022) utilizza un modello linguistico per proporre azioni ad alto livello e una funzione di affordance appresa dall'interazione robotica per valutare la fattibilità delle azioni proposte nell'ambiente concreto. Il grounding si realizza attraverso il collegamento tra le descrizioni linguistiche delle azioni e le rappresentazioni sensorimotore della loro esecuzione. Tuttavia, anche in questo caso, il sistema non costruisce il significato dell'azione dal basso — attraverso un'esperienza incarnata che accumula schemi sensomotori nel senso di Lakoff-Johnson — ma utilizza rappresentazioni pre-apprese di azioni che vengono collegate alla pianificazione linguistica in modo essenzialmente modulare.

La prospettiva più radicalmente integrata è quella proposta da Yoshua Bengio e da altri nella cornice dei **Modelli di Processi Coscienti** (*Conscious Process Models*): modelli che implementano qualcosa di analogo alla Workspace Globale (*Global Workspace Theory*) di Bernard Baars — un'architettura in cui moduli specializzati competono per accedere a uno spazio di lavoro globale condiviso che trasmette l'informazione vincente all'intera architettura. Nella versione di Baars, il workspace globale è il correlato funzionale della coscienza di accesso — il

meccanismo che rende alcune rappresentazioni disponibili per il report verbale, per l'attenzione dirigibile e per il controllo dell'azione. Un modello artificiale che implementasse questa architettura potrebbe, in questa prospettiva, avvicinarsi alla struttura funzionale dell'elaborazione cosciente umana, anche se resta aperta la questione di se una tale implementazione produrrebbe qualcosa di analogo all'esperienza soggettiva.

Queste prospettive di integrazione non sciolgono il problema filosofico ma lo precisano: il problema del grounding non è un problema unico ma una famiglia di problemi distinti che operano a livelli diversi. C'è il problema del grounding percettivo (connettere le rappresentazioni linguistiche a rappresentazioni di qualità sensoriali), il problema del grounding sensomotorio (connettere le rappresentazioni d'azione a sequenze di movimenti incarnati), il problema del grounding emotivo-affettivo (connettere il significato degli stati mentali a stati interni con valenza), e il problema del grounding intersoggettivo (connettere il significato alla comprensione dell'altro come soggetto con stati mentali). I sistemi attuali risolvono parzialmente il primo livello, approssimano il secondo in sistemi robotici, e lasciano sostanzialmente aperti il terzo e il quarto.

IV.7 — Verso una teoria del significato computazionale: convergenze e aporie residue

Le quattro parti di questa monografia hanno percorso, rispettivamente, l'ontologia delle strutture matematiche, la semantica formale del linguaggio, le scienze cognitive del significato e i sistemi artificiali di elaborazione. Ogni parte ha incontrato lo stesso problema da un'angolatura diversa: come fa una struttura — formale, simbolica, neurale, computazionale — a raggiungere ciò che le sta al di là di se stessa? Come fa la forma a toccare il contenuto?

Ciò che emerge dall'esame complessivo non è una risposta ma la mappa di un'aporia strutturale. Il logicismo di Frege ha mostrato che la semantica del linguaggio ordinario richiede un livello di contenuto — il senso — che non si riduce al referente, ma ha incontrato nell'antinomia di Russell il limite di ogni tentativo di derivare il contenuto dalla struttura formale pura. La semantica dei mondi possibili ha fornito un quadro elegante per le modalità e le intensioni, ma ha incontrato il problema dell'iperintensionalità — la distinzione semantica in contesti dove le proposizioni intensionali coincidono — che rimanda a strutture più fini del significato non catturabili dall'approccio estensionale. La grammatica montagoniana ha realizzato il sogno composizionale di derivare il significato di enunciati complessi dai significati delle parti, ma ha lasciato aperto il problema dell'anafora, della quantificazione plurale e dei contesti dipendenti dal discorso.

Le scienze cognitive hanno mostrato che il significato è radicato nell'esperienza incarnata — negli schemi sensomotori, nelle metafore primarie, nella storia evolutiva della cognizione — e che le strutture linguistiche formali galleggiano su questo sfondo corporeo senza poterlo esaurire. La semantica distribuzionale ha catturato la struttura relazionale del significato lessicale con una precisione senza precedenti, ma ha lasciato irrisolto il problema del grounding: le relazioni distribuzionali non si chiudono sul mondo senza passare attraverso le esperienze

dei parlanti che hanno prodotto il corpus. L'intuizionismo di Brouwer ha tentato di fondare il significato matematico sulla costruzione mentale nel tempo interiore, ma ha incontrato il problema dell'intersoggettività: perché le costruzioni di menti diverse convergono sulle stesse strutture?

I sistemi artificiali hanno mostrato che molte delle funzioni cognitive associate alla comprensione del linguaggio — ragionamento contestuale, riferimento, inferenza, narrazione, argomentazione — possono essere implementate in modo convincentemente efficace da sistemi che elaborano sequenze di token secondo principi probabilistici. Questa osservazione ha due possibili letture filosófiche. La prima è deflazionistica: le funzioni cognitive che assomigliamo alla comprensione si riducono, in larga misura, a meccanismi di organizzazione statistica dell'informazione linguistica — e se questi meccanismi sono replicabili artificialmente, forse la nostra concezione del significato era troppo elevata, troppo intessuta di mistero che si rivela non essenziale. La seconda è inflazionistica rispetto ai sistemi artificiali: se questi sistemi raggiungono le funzioni cognitive della comprensione, forse esse sono sufficienti per la comprensione in senso pieno — e il problema del grounding, della coscienza, dell'intenzionalità è una sovrainterpretazione filosofica di un fenomeno che si esaurisce nel funzionale.

Nessuna delle due letture è pienamente soddisfacente. La lettura deflazionistica ignora il dato fenomenico dell'esperienza soggettiva del significato — il fatto che per l'essere umano comprendere non è soltanto eseguire le funzioni corrette ma avere accesso al significato in modo diretto, «da dentro». La lettura inflazionistica attribuisce ai sistemi artificiali qualcosa che le evidenze disponibili non permettono di affermare con certezza: non sappiamo se i modelli di linguaggio di larga scala abbiano qualcosa di analogo all'esperienza soggettiva, e le loro capacità funzionali, per quanto impressionanti, non costituiscono da sole evidenza di coscienza o di intenzionalità originaria.

La posizione che emerge dall'analisi complessiva di questa monografia è quella di un **pluralismo metodologico e ontologico**: non esiste un unico livello di descrizione — né il formale, né il cognitivo, né il computazionale, né il fenomenico — che cattura il significato esaustivamente. Il significato è un fenomeno stratificato che richiede descrizioni complementari a livelli diversi, e ogni riduzione a un solo livello perde qualcosa di essenziale. La semantica formale cattura la struttura composizionale e le relazioni vero-condizionali; la semantica cognitiva cattura il radicamento esperienziale e le strutture prototipiche; la semantica distribuzionale cattura le relazioni di similarità e le strutture latenti nel corpus; la fenomenologia cattura la dimensione soggettiva dell'accesso al contenuto. Nessuno di questi livelli è primo o fondamentale rispetto agli altri: sono aspetti irreducibilmente diversi dello stesso fenomeno.

Questa conclusione può sembrare insoddisfacente dal punto di vista di chi cerca una risposta definitiva. Ma la storia della filosofia suggerisce che le questioni filosófiche autentiche non si chiudono con risposte definitive: si approfondiscono, si precisano, rivelano nuove dimensioni con ogni tentativo di soluzione. Il problema del significato che abita le strutture — che questa monografia ha esaminato nelle sue forme matematiche, linguistiche, cognitive e

artificiali — appartiene a questa specie. La sua irrisolvibilità non è una sconfitta del pensiero ma il segno della sua profondità: è il punto in cui la struttura formale incontra il limite del proprio potere di auto-spiegazione, e in questo incontro rivela qualcosa di essenziale sulla natura sia della struttura sia di ciò che la trascende.

IV.8 — Sistemi neurosimbolici: integrare struttura e apprendimento statistico

Il dibattito tra l'intelligenza artificiale simbolica — che opera attraverso manipolazione esplicita di rappresentazioni strutturate secondo regole formali — e l'intelligenza artificiale subsimbolica — che apprende rappresentazioni distribuite attraverso l'ottimizzazione statistica su esempi — ha attraversato l'intera storia della disciplina, producendo quello che i ricercatori del campo hanno denominato il «grande divario» (*great divide*). La prima generazione di sistemi intelligenti, sviluppata negli anni Cinquanta e Sessanta, era fondamentalmente simbolica: sistemi come il General Problem Solver di Newell e Simon, LISP come linguaggio di rappresentazione della conoscenza, le basi di conoscenza dei sistemi esperti degli anni Ottanta. Il connessionismo degli anni Ottanta e Novanta, seguito dal deep learning del decennio successivo, ha spostato il paradigma dominante verso il subsimbolico. La sfida contemporanea — che costituisce uno dei fronti di ricerca più attivi e filosofica mente interessanti — è quella di integrare i due approcci in architetture ibride che recuperino i rispettivi punti di forza.

I sistemi simbolici tradizionali, che Gary Marcus ha denominato con il termine «intelligenza artificiale di vecchia generazione» (*Good Old-Fashioned AI*, GOFAI), possiedono proprietà che i sistemi subsimbolici hanno faticato a replicare. La **composizionalità** sistematica: un sistema simbolico che rappresenta «Giovanni ama Maria» può sistematicamente rappresentare «Maria ama Giovanni» perché i componenti della rappresentazione sono separabili e ricomponibili. La **trasparenza** delle ragioni: un sistema basato su regole esplicite può rendere conto del proprio ragionamento in termini comprensibili agli esseri umani, articolando la catena di inferenze che ha prodotto un risultato. Il **trasferimento zero-shot**: un sistema simbolico che conosce le regole della logica proposizionale può risolvere qualunque istanza di un problema logico, senza bisogno di aver visto esempi di quella particolare istanza. La **robustezza distribuzionale**: un sistema simbolico non degrada silenziosamente in prossimità dei confini della propria competenza ma segnala esplicitamente quando un problema non rientra nelle regole disponibili.

I sistemi subsimbolici — reti neurali profonde, transformer — possiedono a loro volta proprietà che i sistemi simbolici hanno faticato a replicare. La **percezione**: apprendono rappresentazioni da dati sensoriali grezzi (pixel, forme d'onda) senza richiedere ingegneria delle caratteristiche. La **robustezza al rumore**: gestiscono segnali incompleti, ambigui, parzialmente corrotti in modo naturale. La **generalizzazione statistica**: catturano regolarità statistiche distribuite su enormi corpora che nessuna specifica regola simbolica potrebbe codificare esplicitamente. La **competenza implicita**: acquisiscono abilità linguistiche,

percettive e ragionative attraverso l'esposizione a dati senza richiedere la specifica esplicita delle regole sottostanti.

Paul Smolensky ha proposto, a partire dalle **rappresentazioni tensoriali** studiate nella sezione III.8, l'integrazione più formalmente rigorosa dei due livelli. Il punto chiave è che le reti neurali non sono alternative alle strutture simboliche ma le **implementano** a un livello inferiore di descrizione. Una rappresentazione tensoriale $T = \Sigma_i f_i \otimes r_i$ — dove f_i sono i vettori di riempitivo e r_i i vettori di ruolo — è simultaneamente una struttura discreta (una struttura simbolica con ruoli e riempitivi) e una struttura continua (un vettore in uno spazio di Hilbert). Le operazioni simboliche — binding, unbinding, composizione — si realizzano come operazioni lineari sullo spazio tensoriale. Il vantaggio cruciale di questa architettura ibrida è che permette di apprendere gradualmente sia le strutture di ruolo sia i vettori di riempitivo dai dati, invece di doverli specificare a priori, mantenendo la composizionalità sistematica della rappresentazione simbolica.

La **programmazione differenziabile** (*differentiable programming*) è il paradigma computazionale che ha reso praticabile l'integrazione neuronsimbolica su larga scala. L'idea fondamentale è che qualunque computazione può essere resa differenziabile — dotata di una nozione di gradiente rispetto ai propri parametri — se è implementata come una composizione di operazioni differenziabili. Questo permette di integrare moduli simbolici (parsers sintattici, reasoners logici, database strutturati) all'interno di architetture di deep learning addestrate end-to-end per ottimizzazione del gradiente. Il sistema TensorFlow Fold, i framework PyTorch con autograd, le architetture di reasoning differenziabile come Neural Turing Machines e Differentiable Neural Computers sono esempi di questa integrazione. La Neural Turing Machine di Graves e colleghi (2014) combina un controller neurale — una rete LSTM — con una memoria esterna indirizzabile sia per contenuto sia per posizione, attraverso meccanismi di lettura e scrittura differenziabili. Il sistema apprende a usare la memoria attraverso il gradiente, senza che le operazioni di indirizzamento siano specificate a priori.

La **programmazione probabilistica** rappresenta un'ulteriore dimensione dell'integrazione neuronsimbolica, centrata non sulla struttura sintattica ma sulla struttura generativa. Un programma probabilistico definisce una distribuzione di probabilità su strutture discrete (alberi sintattici, grafici causali, tracce di esecuzione) attraverso un programma generativo che produce campioni dalla distribuzione. Sistemi come Church, WebPPL e Gen permettono di specificare modelli cognitivi — modelli della percezione, del ragionamento, dell'apprendimento — come programmi probabilistici e di condizionare la distribuzione su osservazioni (i dati), inferendo distribuzione posteriore sui parametri latenti attraverso algoritmi di inferenza approssimata (variational inference, sequential Monte Carlo). La inferenza bayesiana nei programmi probabilistici realizza precisamente il tipo di ragionamento strutturato che i sistemi simbolici puri implementano attraverso la logica, ma in un quadro statistico che incorpora l'incertezza e permette l'aggiornamento dalle evidenze.

AlphaFold 2 — il sistema di previsione della struttura delle proteine sviluppato da DeepMind — costituisce un caso paradigmatico di integrazione neuronsimbolica che ha prodotto risultati scientifici di importanza eccezionale. Il

sistema combina reti di attenzione transformer per il processamento delle sequenze aminoacidiche con strutture geometriche equivarianti — reti che rispettano le simmetrie di rotazione e traslazione nello spazio tridimensionale — e con algoritmi di raffinamento iterativo ispirati ai metodi di ottimizzazione delle strutture cristallografiche. La conoscenza biochimica strutturale — espressa nei vincoli geometrici sugli angoli di torsione, nelle regole di Ramachandran — è integrata nell'architettura non come post-processing esterno ma come vincolo strutturale nell'architettura della rete. Il risultato è un sistema che apprende dal corpus delle strutture proteiche note ma incorpora la struttura simbolica della biologia strutturale come vincolo architetturale.

I modelli di linguaggio di larga scala hanno recentemente sviluppato la capacità di interagire con strumenti simbolici esterni — calcolatrici, interpreti di codice, motori di ricerca, basi di dati strutturate — attraverso i meccanismi di «uso degli strumenti» (*tool use*). Questo crea un sistema ibrido di tipo diverso da quelli discussi sopra: il componente neurale gestisce la comprensione del linguaggio naturale, la pianificazione e la generazione di output, mentre i componenti simbolici gestiscono la computazione esatta, la ricerca nelle basi di dati e l'esecuzione di codice. La produzione di soluzioni corrette a problemi matematici, la generazione di codice funzionante, la ricerca di informazioni aggiornate dipendono dal componente simbolico, mentre la comprensione contestuale e la pianificazione linguistica dipendono dal componente neurale. Questo sistema ibrido supera i limiti di entrambi i componenti presi separatamente, realizzando in modo pratico l'integrazione che la ricerca neuronsimbolica teorica perseguiva.

La domanda filosofica che i sistemi neuronsimbolici pongono concerne la natura del significato nelle architetture ibride. Il significato nei sistemi simbolici classici è derivato dall'interpretazione: i simboli significano ciò che l'interprete decide che significhino, e questa relazione interpretativa non è interna al sistema ma è stabilita dall'esterno. Il significato nei sistemi subsimbolici è distribuzionale: emerge dalle strutture statistiche delle rappresentazioni vettoriali. Il significato nei sistemi ibridi è stratificato: il componente simbolico porta un tipo di contenuto strutturale e composizionale, il componente neurale porta un tipo di contenuto distribuzionale e contestuale, e l'integrazione tra i due produce emergenza di contenuto a un livello superiore che nessuno dei due componenti possiederebbe separatamente. Questa stratificazione del significato nei sistemi ibridi rispecchia la stratificazione che le scienze cognitive e la filosofia del linguaggio hanno identificato nel significato umano: la struttura composizionale delle rappresentazioni linguistiche, il radicamento distribuzionale nel corpus dell'esperienza, e la connessione pragmatica con le pratiche d'uso e le forze illocutorie degli atti. Il sistema ibrido neurosimbolico è, in questo senso, una metafora computazionale della complessità filosofica del significato — non la sua soluzione, ma la sua rispecchiamento strutturale.

CONCLUSIONE: IL SIGNIFICATO COME LIMITE DELLA STRUTTURA

La geometria dell'irrisolto

Una monografia che percorra quattro regioni concettuali — l'ontologia matematica, la filosofia del linguaggio, le scienze cognitive, l'intelligenza artificiale — rischia di risultare, al termine del percorso, o troppo dispersa per avere un filo conduttore riconoscibile o troppo convergente per rendere giustizia alla complessità autentica dei problemi. Questo scritto ha scelto deliberatamente di non sacrificare la complessità alla coerenza: il filo conduttore non è una tesi unica che si consolida attraverso le parti, ma un problema che si ripropone — invariato nella sua struttura profonda, variato nelle sue manifestazioni tecniche — ad ogni livello di esame. Il problema è quello del significato come proprietà relazionale tra struttura e mondo: come fa una struttura — un sistema di assiomi, una grammatica, un corpus di testo, una rete neurale — a raggiungere ciò che le è esterno?

L'analisi condotta nelle quattro parti permette ora di identificare con maggiore precisione la struttura dell'aporia. Essa si manifesta, in ognuna delle quattro regioni, come una tensione tra due polarità che resistono alla riduzione reciproca.

Nell'ontologia matematica, la tensione è tra la **struttura formale** e l'**oggetto matematico**. Il formalismo — nella versione hilbertiana o nella versione assiomatica di ZFC — tratta la matematica come lo studio di sistemi di simboli soggetti a regole di trasformazione, senza impegno ontologico rispetto agli oggetti che tali sistemi descrivono. Il platonismo — nelle sue versioni gödelliana, fregeana, neo-fregeana — afferma che gli oggetti matematici esistono indipendentemente dai sistemi formali che li descrivono, e che i sistemi formali sono strumenti per esplorare una realtà che li trascende. Lo strutturalismo — nelle versioni di Shapiro, Resnik, Hellman — tenta una posizione intermedia: gli oggetti matematici non sono né entità platoniche autonome né costruzioni formali arbitrarie, ma posizioni in strutture, e le strutture sono determinate dai loro isomorfismi. Nessuna di queste posizioni dissolve la tensione; ognuna la sposta, la precisamente, la rende più visibile. Il Lemma di Yoneda cattura il massimo di questo strutturalismo radicalizzandolo categorialmente: un oggetto è interamente determinato dalle sue relazioni morfiche con tutti gli altri oggetti del suo universo. Ma ciò che le relazioni morfiche determinano è la posizione nella struttura, non il contenuto «da dentro» — la stessa aporia del significato riproposta al livello categoriale.

Nella filosofia del linguaggio, la tensione è tra il **senso** e la **denotazione**, tra il **modo di presentazione** e il **referente**. Frege ha mostrato che il referente non esaurisce il significato: due espressioni co-denotanti possono differire per contenuto cognitivo, e questa differenza — il senso — è irriducibile alla denotazione. La teoria di Kripke ha mostrato che i nomi propri sono designatori rigidi che denotano il medesimo individuo in tutti i mondi possibili, aprendo la

categoria delle verità necessarie a posteriori e rivelando che la mappa concettuale della tradizione analitica era inadeguata. La semantica dei mondi possibili ha fornito un quadro elegante per le intensioni, ma ha incontrato l'iperintensionalità — la distinzione semantica in contesti in cui le intensioni coincidono — che mostra come nemmeno le proposizioni intensionali esauriscano il significato. La grammatica montagoniana ha realizzato il Principio di Composizionalità in forma matematica precisa, ma ha lasciato insoluti i fenomeni che richiedono la dimensione dinamica del discorso, la forza illocutoria degli atti linguistici, il grounding corporeo dei contenuti astratti. Il significato del linguaggio si rivela stratificato — condizioni di verità, forza illocutoria, contenuto cognitivo, radicamento esperienziale, struttura distribuzionale — e nessuno strato si lascia ridurre agli altri.

Nelle scienze cognitive, la tensione è tra la **struttura simbolica** della cognizione e la sua **radice incarnata**. Il cognitivismo computazionale — nella versione della LOTH di Fodor — afferma che i processi cognitivi sono computazioni su rappresentazioni simboliche strutturate, e che la struttura sintattica delle rappresentazioni è il meccanismo che permette alle inferenze di rispettare le relazioni semantiche tra i contenuti. Il connessionismo afferma che i processi cognitivi sono pattern di attivazione distribuiti su reti di unità semplici, senza struttura simbolica localizzata, e che la sistematicità e la produttività emergono dalla struttura globale della rete piuttosto che dalla struttura locale di entità simboliche discrete. L'enattivismo afferma che la cognizione non è né computazione simbolica né propagazione di attivazioni neurali, ma accoppiamento sensomotorio tra organismo e ambiente, e che il significato emerge da questo accoppiamento — non è né nelle rappresentazioni interne né negli oggetti del mondo ma nella relazione dinamica tra i due. Queste tre posizioni non sono semplicemente incompatibili: catturano aspetti diversi di un fenomeno multi-livello, e la domanda di quale sia il livello «corretto» di descrizione della cognizione non ha una risposta unica e definitiva.

Nell'intelligenza artificiale, la tensione è tra la **competenza comportamentale** e la **comprensione semantica**. I modelli di linguaggio su larga scala raggiungono prestazioni che sembrano richiedere qualcosa di analogo alla comprensione — risolvono problemi complessi, ragionano per analogia, producono discorsi coerenti e contestualmente appropriati, identificano relazioni strutturali sottili tra concetti distanti. Ma la questione di se questa competenza comportamentale implichi comprensione nel senso filosofica mente rilevante — intenzionalità genuina, accesso al significato dall'interno, capacità di rispondere alla forza illocutoria degli atti e non soltanto al contenuto proposizionale — resta aperta. L'argomento di Searle, anche se non conclusivo, pone un punto che non si può liquidare facilmente: la manipolazione formale di simboli, per quanto sofisticata e per quanto appropriata nelle sue produzioni comportamentali, non è evidenza sufficiente dell'esistenza di contenuto semantico genuino.

La struttura trasversale dell'aporia

Questo esame rivela una struttura trasversale che attraversa le quattro regioni: il significato è sempre in eccedenza rispetto alla struttura che lo porta. Il senso

fregeano eccede il referente; il contenuto cognitivo eccede le condizioni di verità; lo schema sensomotorio eccede la sua codifica formale; il noema husserliano eccede il sistema di simboli che lo esprime; il qualia eccede la descrizione funzionale. Ogni tentativo di catturare esaustivamente il significato in una struttura formale produce una struttura più ricca che cattura qualcosa di autentico, e al tempo stesso rivela qualcosa che eccede la cattura.

Questo non è misticismo né irrazionalismo. È la struttura profonda di un problema che resiste alla riduzione non perché non sia rigoroso abbastanza da lasciarsi formalizzare, ma perché ogni formalizzazione produce un sistema che a sua volta richiede un'interpretazione — un atto di comprensione che colloca il sistema nel mondo e attribuisce ai suoi simboli un contenuto. Il contenuto non è dentro il sistema; è la relazione tra il sistema e ciò che è al di fuori di esso. Ma questa relazione non può essere catturata interamente dall'interno del sistema, perché richiederebbe di descrivere la relazione tra il sistema e il mondo usando risorse del sistema stesso — e questa descrizione sarebbe, di nuovo, un sistema che richiede interpretazione. Il circolo non è vizioso nel senso logico — non produce contraddizioni — ma è strutturalmente inesauribile: ogni livello di meta-descrizione rinvia a un livello superiore.

Questa struttura è precisamente quella che Gödel ha rivelato nella matematica: nessun sistema formale sufficientemente potente può dimostrare la propria coerenza, perché la coerenza è una proprietà che il sistema ha rispetto al mondo esterno (ai modelli del sistema), e non è interna al sistema stesso. I teoremi di incompletezza non dicono che la verità matematica trascende la dimostrazione in un senso platonico misterioso; dicono che ogni sistema formale ha una prospettiva che non include la propria posizione rispetto al mondo — che il significato del sistema non è interamente disponibile dall'interno del sistema. Questa è la versione matematicamente precisa dell'aporia filosofica del significato.

Il significato come orizzonte, non come dato

La conclusione che emerge non è nichilistica né scettica. Non afferma che il significato non esiste o che la comprensione è un'illusione. Afferma qualcosa di più preciso e di più interessante: il significato non è un dato ma un **orizzonte** — una struttura che orienta e guida il pensiero senza mai essere interamente disponibile come oggetto. Gli esseri umani abitano l'orizzonte del significato in modo pratico, pre-teorico, prima di qualsiasi formalizzazione: parlano, comprendono, rispondono alla forza illocutoria degli atti, interpretano gli stati mentali degli altri, costruiscono e abitano mondi simbolici condivisi. Questa competenza pratica è ciò che le scienze cognitive tentano di descrivere, la filosofia del linguaggio di formalizzare, la matematica di astrarre, i sistemi artificiali di simulare.

Nessuna di queste imprese è futile: ognuna rivela struttura autentica, cattura aspetti reali del fenomeno, produce conoscenza genuina. Ma ognuna opera dall'interno di una prospettiva che non include la propria posizione rispetto all'orizzonte del significato: il matematico che costruisce la teoria delle categorie opera già nell'orizzonte del significato quando attribuisce contenuto ai morfismi;

il semanticista che costruisce il modello montagoniano opera già nell'orizzonte del significato quando interpreta le condizioni di verità come vere o false rispetto a una struttura; il modello di linguaggio che produce output linguisticamente appropriato opera già nell'orizzonte del significato quando viene interpretato come portatore di contenuto dai suoi utenti umani.

Il problema del significato — nella sua forma più profonda — è il problema del rapporto tra una prospettiva e l'orizzonte che la rende possibile senza essere interamente visibile da essa. Questo non è un problema che si risolve trovando la teoria giusta; è un problema che si porta avanti approfondendolo — precisando le strutture formali che lo articolano, esplorando le funzioni cognitive che lo incarnano, costruendo i sistemi artificiali che lo approssimano, e riconoscendo in ogni approssimazione ciò che resta oltre la portata dell'approssimazione stessa.

La matematica, il linguaggio, la cognizione e i sistemi artificiali sono quattro modalità di questo portare avanti — quattro modi in cui la struttura tende verso il significato che la trascende. Il loro studio congiunto non produce una risposta definitiva alla domanda che ha aperto questa monografia, ma produce qualcosa di più prezioso: la comprensione precisa di perché la domanda non si chiude, e la mappa dei luoghi in cui questa apertura si fa più visibile e più produttiva.

APPENDICE I: Teoria dei Tipi Dipendenti come Fondamento della Prova Formale Assistita

Dalle teorie dei tipi semplici alla dipendenza: genealogia di un'idea

La teoria dei tipi nasce, nella sua versione moderna, come risposta tecnica alla crisi dei fondamenti provocata dalla scoperta dei paradossi negli insiemi naivi. Russell, nella stessa lettera che comunicava a Frege la contraddizione nell'impianto dei *Grundgesetze*, stava elaborando la propria soluzione: introdurre una gerarchia di tipi che impedisse ai predicati di applicarsi a se stessi o ad elementi dello stesso livello. La *Principia Mathematica* di Whitehead e Russell (1910–1913) presentava una teoria ramificata dei tipi in cui le proposizioni erano organizzate in ordini e la regola di stratificazione impediva la formazione di proposizioni autoreferenziali. Il sistema era tecnicamente rigoroso ma filosofica mente oneroso: l'assioma di riducibilità (*axiom of reducibility*), introdotto per recuperare la matematica classica all'interno della gerarchia ramificata, era apertamente ad hoc e privava l'intera costruzione della trasparenza filosofica che avrebbe dovuto avere.

La semplificazione fondamentale fu compiuta da Frank Ramsey nel 1925 e adottata poi da Church: la teoria dei tipi semplici (*simple type theory*, STT) elimina la ramificazione e conserva soltanto la stratificazione per ordini. In STT ogni espressione ha un tipo unico; i tipi di base sono gli individui ι e le proposizioni o; per ogni coppia di tipi σ e τ esiste il tipo funzionale $\sigma \to \tau$ delle funzioni da σ a τ.

La regola di tipizzazione fondamentale è quella dell'applicazione: se f ha tipo $\sigma \to \tau$ e a ha tipo σ, allora f(a) ha tipo τ. La formazione di paradossi è bloccata perché un predicato di tipo $\iota \to o$ non può applicarsi a se stesso, che ha tipo $\iota \to o$, poiché $\iota \to o \neq \iota$.

La **corrispondenza di Curry-Howard** — il risultato che stabilisce un isomorfismo strutturale tra le prove in certi sistemi di deduzione naturale e i termini del lambda-calcolo tipizzato — ha trasformato la teoria dei tipi da tecnica fondazionale in strumento computazionale e filosofico di straordinaria portata. L'isomorfismo si articola nel modo seguente: ogni tipo A corrisponde a una proposizione, ogni termine t di tipo A corrisponde a una prova della proposizione A, la costruzione di un termine di tipo A mediante le regole del lambda-calcolo tipizzato corrisponde alla costruzione di una prova di A mediante le regole della deduzione naturale. Le regole logiche fondamentali hanno la seguente corrispondenza:

Introduzione dell'implicazione (astrazione lambda): se $\Gamma, x{:}A \vdash t{:}B$ allora $\Gamma \vdash \lambda x.t : A \to B$.

Eliminazione dell'implicazione (applicazione): se $\Gamma \vdash f{:}A \to B$ e $\Gamma \vdash a{:}A$ allora $\Gamma \vdash f(a){:}B$.

Introduzione della congiunzione (accoppiamento): se $\Gamma \vdash a{:}A$ e $\Gamma \vdash b{:}B$ allora $\Gamma \vdash \langle a,b \rangle : A \times B$.

Eliminazione della congiunzione (proiezione): se $\Gamma \vdash p{:}A \times B$ allora $\Gamma \vdash \pi_1(p){:}A$ e $\Gamma \vdash \pi_2(p){:}B$.

Nella prospettiva della corrispondenza, un programma che termina con un valore di tipo A è una costruzione effettiva della proposizione A: la computazione è dimostrazione. La riduzione di un termine — la sua valutazione secondo le regole di β-riduzione del lambda-calcolo, $(\lambda x.t)(a) \rightsquigarrow t[x:=a]$ — corrisponde alla normalizzazione di una prova: l'eliminazione dei «detours» dimostrativi, ossia le sequenze di introduzione e immediata eliminazione di un connettivo.

La teoria dei tipi semplici, tuttavia, non è abbastanza espressiva per catturare la matematica ordinaria. Le proprietà interessanti degli oggetti matematici dipendono dai valori di certi termini: il tipo dei vettori in $\mathbb{R}^n$ dipende dal numero naturale n, il tipo delle matrici invertibili dipende dalla proposizione che il determinante è non nullo, il tipo delle liste di lunghezza esatta k dipende dal numero k. Per rappresentare queste proprietà in modo che la loro correttezza sia verificabile al momento della tipizzazione — prima dell'esecuzione — è necessario introdurre tipi che dipendono da termini: i **tipi dipendenti**.

La struttura formale dei tipi dipendenti: tipi Π, Σ e il tipo identità

La teoria dei tipi dipendenti nella sua forma matura fu elaborata da Per Martin-Löf in una serie di lavori tra il 1971 e il 1984, culminati nel volume *Intuitionistic Type Theory* (1984). Il sistema di Martin-Löf — noto come MLTT — è

costruttivista: tutta la matematica vi è sviluppata a partire da costruzioni effettive, ogni oggetto matematico è un termine computazionale e la prova dell'esistenza di un oggetto deve fornire il metodo per costruirlo. Ma la ricchezza espressiva del sistema, garantita dalla presenza dei tipi dipendenti, permette di sviluppare l'intera matematica intuizionista e una porzione considerevole di quella classica.

Il tipo dipendente fondamentale è il **tipo Π** (prodotto dipendente o tipo funzione dipendente). Dato un tipo A e una famiglia di tipi B dipendente da termini di tipo A — una funzione $B : A \to \mathcal{U}$ che assegna a ogni a:A un tipo B(a) nell'universo $\mathcal{U}$ — il tipo Π(x:A).B(x) è il tipo delle funzioni che, dato un termine a:A, producono un termine di tipo B(a). La dipendenza è nel codominio: il tipo del risultato varia al variare dell'argomento.

Regola di formazione:

$$\mathbf{A : \mathcal{U},\ x{:}A \vdash B(x) : \mathcal{U}}$$

$$\Pi(x{:}A).B(x) : \mathcal{U}$$

Regola di introduzione (astrazione dipendente):

$$x{:}A \vdash b(x) : B(x) \quad \text{——————————————} \quad \lambda x.b(x) : \Pi(x{:}A).B(x)$$

Regola di eliminazione (applicazione dipendente):

$$\mathbf{f : \Pi(x{:}A).B(x),\ a : A}$$

$$f(a) : B(a)$$

con regola di computazione f(a) ≡ b(a) (β-riduzione). Quando B non dipende da x, il tipo Π(x:A).B si riduce al tipo funzione ordinario A→B: il tipo Π è una generalizzazione del tipo funzionale che cattura le funzioni i cui risultati hanno tipi che variano al variare degli argomenti.

Nella corrispondenza di Curry-Howard, il tipo Π(x:A).B(x) corrisponde alla proposizione $\forall$ x:A.B(x): un termine di tipo Π(x:A).B(x) è una funzione che, dato un qualsiasi a:A, produce una prova di B(a), che è precisamente ciò che costituisce una prova del quantificatore universale nella concezione intuizionista di Brouwer-Heyting-Kolmogorov (BHK).

Il **tipo Σ** (somma dipendente o coppia dipendente) ha struttura duale. Dati $A : \mathcal{U}$ e $B : A \to \mathcal{U}$, il tipo Σ(x:A).B(x) è il tipo delle coppie dipendenti ⟨a,b⟩ dove a:A e b:B(a): il tipo del secondo componente dipende dal valore del primo.

Regola di formazione:

$$\mathbf{A : \mathcal{U},\ x{:}A \vdash B(x) : \mathcal{U}}$$

$$\Sigma(x{:}A).B(x) : \mathcal{U}$$

Regola di introduzione:

$$a : A,\ b : B(a) \quad \text{——————————————} \quad \langle a,b \rangle : \Sigma(x:A).B(x)$$

Regole di eliminazione (proiezioni):

$$p : \Sigma(x:A).B(x) \quad \text{——————————————} \quad \pi_1(p) : A$$

$$p : \Sigma(x:A).B(x) \quad \text{——————————————} \quad \pi_2(p) : B(\pi_1(p))$$

Nella corrispondenza BHK, il tipo $\Sigma(x:A).B(x)$ corrisponde alla proposizione $\exists\, x:A.B(x)$: una prova dell'esistenziale è una coppia costituita da un testimone a:A e da una prova b che a soddisfa la proprietà B. Quando B non dipende da x, $\Sigma(x:A).B$ si riduce al prodotto cartesiano $A \times B$.

Il **tipo identità** — o tipo di uguaglianza proposizionale — è la terza costruzione fondamentale di MLTT. Dati $A:\mathcal{U}$ e a,b:A, il tipo Id_A(a,b) (scritto anche a =_A b o semplicemente a = b quando il tipo è chiaro dal contesto) è il tipo delle prove che a e b sono identici in A.

Regola di introduzione (riflessività):

$$a : A \quad \text{——————————} \quad \text{refl_a} : \text{Id_A}(a,a)$$

La regola di eliminazione — il **principio J**, o *path induction* — afferma che, per dimostrare una proprietà C di tutti i termini di tipo Id_A(a,b) per tutti gli a,b:A, è sufficiente dimostrarla per refl_a per ogni a:A:

$$\frac{C : \Pi(x:A).\Pi(y:A).\ \text{Id_A}(x,y) \to \mathcal{U} \qquad d : \Pi(x:A).\ C(x, x, \text{refl_x}) \qquad p : \text{Id_A}(a,b)}{J(C, d, a, b, p) : C(a, b, p)}$$

La regola J asserisce che la struttura delle prove di identità è completamente determinata dalla riflessività: ogni prova di identità si comporta, per tutti gli scopi dimostrativo-computazionali, come una riflessività. Questo principio è la radice della connessione tra tipi dipendenti e topologia algebrica nella teoria dell'omotopia dei tipi.

La potenza espressiva congiunta di Π, Σ e del tipo identità permette di codificare l'intera matematica di ordine superiore come termine computazionale. Il tipo delle funzioni iniettive da A a B è $\Sigma(f:A\to B).\Pi(x:A).\Pi(y:A).\text{Id_B}(f(x), f(y)) \to \text{Id_A}(x,y)$; il tipo delle funzioni biettive è $\Sigma(f:A\to B).\Sigma(g:B\to A).\ [\Pi(x:A).\text{Id_A}(g(f(x)),x)] \times [\Pi(y:B).\text{Id_B}(f(g(y)),y)]$; il tipo dei numeri naturali è definito induttivamente con i costruttori zero : $\mathbb{N}$ e succ : $\mathbb{N}\to\mathbb{N}$, e ogni enunciato sull'aritmetica si traduce in un tipo la cui abitazione corrisponde alla verità dell'enunciato.

Gli assistenti di prova: Coq, Agda, Lean e la verifica formale assistita

La teoria dei tipi dipendenti è diventata strumento operativo attraverso la sua implementazione in sistemi software noti come **assistenti di prova** o **dimostratori interattivi di teoremi** (*proof assistants*, *interactive theorem provers*). Questi sistemi permettono all'utente di costruire prove matematiche formali interattivamente, con la garanzia che ogni passo sia verificato meccanicamente: il kernel del sistema controlla che ogni applicazione di regola sia legittima, e la correttezza di una prova completa è garantita dalla correttezza del kernel, che è tipicamente un programma di piccole dimensioni e alta verificabilità.

Il sistema **Coq** — sviluppato presso l'INRIA a partire dai lavori di Thierry Coquand e Gérard Huet negli anni Ottanta, ancora attivamente mantenuto — è fondato sul **Calcolo delle Costruzioni Induttive** (CCI o CIC). Il CIC stratifica i tipi in due universi distinti: **Prop**, l'universo delle proposizioni logiche, i cui termini sono prove e che gode della *proof irrelevance* (tutte le prove della stessa proposizione sono identificate ai fini della computazione), e **Type_i** (per i = 0, 1, 2, ...), la gerarchia dei tipi computazionali. I tipi induttivi in Coq sono definiti attraverso una lista di costruttori; le funzioni su tipi induttivi sono definite per ricorsione strutturale, la cui terminazione è verificata staticamente attraverso un controllo di decrescita strutturale degli argomenti.

Per i numeri naturali, la definizione induttiva in Coq è:

Inductive nat : Type := | O : nat | S : nat → nat.

La funzione di addizione è:

Fixpoint add (n m : nat) : nat := match n with | O => m | S n' => S (add n' m) end.

Il sistema verifica che add decresce strutturalmente sul primo argomento n (da S n' a n' ad ogni passo ricorsivo), garantendo la terminazione e dunque la consistenza logica del sistema. Un sistema che ammettesse ricorsioni non terminanti sarebbe logicamente triviale: si potrebbe costruire un termine di qualsiasi tipo per mezzo di una ricorsione che non termina mai, dimostrando proposizioni false.

Tra le realizzazioni più significative della verifica formale in Coq vi è la **dimostrazione del Teorema dei Quattro Colori** da parte di Georges Gonthier e Benjamin Werner (2005). Il teorema afferma che ogni mappa planare è colorabile con al più quattro colori in modo che regioni adiacenti abbiano colori diversi; la sua prova originale di Appel e Haken (1977) aveva richiesto la verifica computerizzata di 1482 configurazioni di grafi ridotti, sollevando domande sull'affidabilità della verifica computerizzata non certificata. La prova di Gonthier ha formalizzato l'intera catena di ragionamento — inclusa la verifica delle configurazioni — in modo che la correttezza sia garantita dal kernel di Coq senza fare affidamento su programmi esterni non verificati. Una realizzazione ancora più ambiziosa è la **prova del Teorema di Feit-Thompson** — l'enunciato che

ogni gruppo di ordine dispari è risolubile, la cui dimostrazione originale di Feit e Thompson del 1963 occupava 255 pagine — completata da Georges Gonthier e un team di collaboratori in Coq nel 2012, per un totale di circa 170.000 righe di codice formale.

Il sistema **Lean** — sviluppato da Leonardo de Moura a partire dal 2013, nella versione Lean 4 co-sviluppata con Sebastian Ullrich — ha acquisito rilievo crescente attraverso il progetto **Mathlib**, una libreria comunitaria che nel 2024 superava il milione di righe di codice formale comprendente migliaia di teoremi della matematica universitaria e avanzata. Lean 4 adotta una variante della teoria dei tipi dipendenti nota come **Calcolo delle Costruzioni con Universi** con tipi quoziente proposizionali. Una peculiarità di Lean 4 è l'integrazione tra il linguaggio delle prove e un linguaggio di programmazione general-purpose: le stesse costruzioni linguistiche servono sia per scrivere algoritmi efficienti sia per costruire prove matematiche, realizzando in modo particolarmente diretto l'identificazione proposizioni-tipi della corrispondenza Curry-Howard.

Il sistema **Agda** — sviluppato da Ulf Norell sulla base di lavori di Catarina Coquand — è strutturalmente più vicino alla teoria di Martin-Löf e non stratifica gli universi in Prop e Type: tutte le proposizioni sono tipi, e le prove sono termini come qualsiasi altro. Questa omissione della stratificazione rende Agda particolarmente adatta alla formalizzazione di risultati in cui la struttura delle prove è matematicamente rilevante, come avviene nella HoTT, dove le prove di uguaglianza sono percorsi omotopici e la loro struttura è l'oggetto di studio.

La Teoria dell'Omotopia dei Tipi: percorsi, univalenza e i fondamenti strutturali

La **Teoria dell'Omotopia dei Tipi** (HoTT) è la sintesi tra la teoria dei tipi dipendenti di Martin-Löf e la topologia algebrica moderna, emersa da lavori di Vladimir Voevodsky, Steve Awodey, Michael Shulman e altri nel primo decennio del XXI secolo e sistematizzata nel volume collettivo *Homotopy Type Theory: Univalent Foundations of Mathematics* (2013). L'osservazione fondamentale è che il tipo identità Id_A(a,b) di Martin-Löf si comporta formalmente come lo spazio dei **percorsi** (*paths*) tra i punti a e b in uno spazio topologico.

Nella topologia algebrica, un percorso da a a b in uno spazio X è una funzione continua p:[0,1]→X con p(0)=a e p(1)=b; due percorsi sono **omotopi** se uno può essere continuamente deformato nell'altro mantenendo fissi gli estremi. La composizione di percorsi definisce un'operazione di gruppo sulle classi di omotopia; il **gruppo fondamentale** $\pi_1(X,a)$ di uno spazio è il gruppo delle classi di omotopia dei percorsi chiusi basati in a. L'interpretazione omotopica della teoria dei tipi assegna a ogni tipo A il ruolo di uno spazio topologico, a ogni termine a:A il ruolo di un punto, e a ogni prova p:Id_A(a,b) il ruolo di un percorso da a a b. Le prove di uguaglianza tra prove di uguaglianza — i termini di tipo Id_{Id_A(a,b)}(p,q) — sono omotopie tra percorsi; i termini di livello superiore sono omotopie di omotopie, e così via in una gerarchia potenzialmente infinita. I tipi hanno la struttura di ∞**-groupoid**: oggetti, morfismi, 2-morfismi, ... tutti invertibili.

Il **Lemma di univalenza**, congetturato da Voevodsky e assunto come assioma fondamentale in HoTT, asserisce:

$(A \simeq B) \simeq (A =_{\mathcal{U}} B)$

dove $A \simeq B$ è il tipo delle **equivalenze** tra A e B — le funzioni $f:A\to B$ che hanno un'inversa omotopica, ovvero una funzione $g:B\to A$ tale che $\Pi(x:A).Id_A(g(f(x)),x)$ e $\Pi(y:B).Id_B(f(g(y)),y)$ — e $A =_{\mathcal{U}} B$ è il tipo identità tra A e B nell'universo $\mathcal{U}$. L'assioma afferma che la funzione naturale

$idtoeqv : (A =_{\mathcal{U}} B) \to (A \simeq B)$

— che manda $refl_A$ nell'identità $id_A : A \simeq A$ — è essa stessa un'equivalenza, e ha dunque un'inversa

$ua : (A \simeq B) \to (A =_{\mathcal{U}} B)$

soddisfacente la condizione di computazione: per ogni $f : A \simeq B$ e ogni x:A,

$transport\ P\ (ua\ f)\ x \equiv f(x)$

dove $transport\ P\ p : P(a) \to P(b)$ è la funzione di trasporto lungo il percorso $p:a =_A b$ per una famiglia $P:A\to\mathcal{U}$.

L'assioma di univalenza ha una conseguenza immediatamente comprensibile dal punto di vista matematico: le **proprietà matematicamente esprimibili** — le proprietà che si possono formulare nel linguaggio della teoria dei tipi — sono invarianti per equivalenza. Se $A \simeq B$ e $P:\mathcal{U}\to\mathcal{U}$ è una proprietà, allora $P(A) \simeq P(B)$. Questo è il principio di **invarianza per equivalenza** (*equivalence invariance* o *transport principle*): strutture equivalenti sono indistinguibili per qualsiasi proprietà tipizzabile. In ZFC, due insiemi possono essere isomorfi ma distinti, e si possono formulare proprietà che li distinguono (per esempio, la proprietà di contenere ∅ come elemento). In HoTT con univalenza, i tipi equivalenti sono identici nell'universo, e nessuna proprietà tipizzabile può distinguerli.

La conseguenza fondazionale è che in HoTT la consuetudine matematica di lavorare «a meno di isomorfismo» — di trattare strutture isomorfe come intercambiabili — diventa un teorema invece di una convenzione informale. Il principio di Leibniz dell'indiscernibilità degli identici — se $a = b$ allora $P(a) \leftrightarrow P(b)$, il *Principio di Sostitutività* — vale in tutte le teorie dei tipi; in HoTT con univalenza vale anche una forma del suo converso per le proprietà matematicamente esprimibili.

La stratificazione interna dei tipi per **livello omotopico** (*h-level*) fornisce una classificazione fine della complessità della struttura di identità di un tipo. Un tipo A è **contrattibile** (h-level -2) se ha esattamente un elemento a meno di uguaglianza: $isContr(A) := \Sigma(a:A).\Pi(x:A).Id_A(a,x)$. È una **proposizione** (h-level -1) se tutti i suoi elementi sono uguali: $isProp(A) := \Pi(x:A).\Pi(y:A).Id_A(x,y)$. È un **insieme** (h-level 0) se tutte le sue prove di uguaglianza sono proposizioni: $isSet(A) := \Pi(x:A).\Pi(y:A).isProp(Id_A(x,y))$. Il livello 1 corrisponde ai **groupoid**, il livello 2 alle **bigruppoidi**, e così via. La gerarchia permette di precisare in quale senso i tipi «si comportano come insiemi» (quando sono a livello 0) o «hanno

struttura topologica non banale» (quando sono a livello ≥ 1 per la presenza di percorsi non banali tra i loro elementi).

Verifica formale e struttura epistemica della certezza matematica

La domanda filosofica che i sistemi di verifica formale sollevano concerne la natura della certezza matematica che essi pretendono di fornire. Una prova in Lean o in Coq è primariamente una computazione che termina senza errori nel kernel del sistema: la sua correttezza è garantita non da un argomento che persuade un matematico competente, ma dal controllo meccanico di un programma. L'affidabilità della prova dipende dunque dall'affidabilità del kernel: se quest'ultimo contenesse un bug che rende accettabili certe derivazioni di enunciati falsi, le prove prodotte dal sistema non garantirebbero la verità degli enunciati dimostrati.

Questo è il **problema della fiducia nel compilatore** (*trusting the compiler problem*), che Ken Thompson aveva formulato nel 1984 a proposito della catena compilatore-sistema operativo ma che si applica con altrettanta precisione ai sistemi di verifica formale. La regresso si ferma non perché si raggiunga un fondamento assolutamente certo, ma perché si decide di fidarsi di uno strato specifico della gerarchia computazionale. La risposta pratica è il **de Bruijn criterion**: il nucleo del sistema deve essere sufficientemente piccolo da poter essere verificato manualmente da un esperto. I kernel di Coq e Lean 4 ammontano a poche migliaia di righe di codice — un'entità verificabile con sforzo ragionevole, a differenza dell'enorme mole di codice dei sistemi di tattica costruiti sopra di essi.

Una dimensione filosofica più sottile riguarda il rapporto tra verifica e comprensione. Una prova formale di un teorema complesso è un termine nel calcolo dei tipi che il sistema ha verificato essere di tipo corretto; ma questo termine può essere opaco, prodotto da tattiche automatiche che costruiscono prove senza che l'utente le comprenda. Emerge un rischio di dissociazione: la certezza è garantita meccanicamente, ma la comprensione — la risposta alla domanda *perché* il teorema è vero — non è assicurata dalla verifica. I teorema dimostrati possono essere certificati senza essere capiti. Il progetto di formalizzazione in Lean della **congettura di Fermat-Wiles** attraverso la dimostrazione di Wiles del 1995 — che mobilita teorie vastissime come la geometria algebrica, la teoria delle rappresentazioni e la geometria di Shimura-Taniyama — renderà visibile questa tensione nella sua forma più acuta: la prova formale, qualora completata, sarà certa ma la sua comprensione richiederà una ricostruzione intellettuale che il termine di prova da solo non fornisce.

Il confronto con il **problema del frame** è illuminante: entrambi rivelano la stessa struttura di fondo. Il frame problem emerge dalla difficoltà di esplicitare la conoscenza di sfondo implicita nelle competenze degli agenti intelligenti; la verifica formale richiede l'esplicitazione di ogni passo di ragionamento che un matematico darebbe per scontato. In entrambi i casi, la formalizzazione è

possibile ma costosa: si può catturare il ragionamento esplicito a prezzo di perdere ciò che lo sostiene tacitamente, sia nelle pratiche di pianificazione degli agenti sia nelle pratiche dimostrative dei matematici.

APPENDICE II: La Crisi della Riproducibilità nell'Intelligenza Artificiale come Questione Epistemica

Anatomia di una crisi: dalla scienza empirica all'apprendimento automatico

La crisi della riproducibilità — la scoperta, nel secondo decennio del XXI secolo, che una quota significativa dei risultati pubblicati nelle scienze empiriche non sopravviveva al tentativo di replicazione sistematica da parte di team indipendenti — è stata dapprima identificata e dibattuta nella psicologia sperimentale, attraverso il celebre *Reproducibility Project* (Open Science Collaboration, 2015), che riuscì a replicare con effetti analoghi soltanto il 36–39% di un campione di 100 studi pubblicati in riviste di alto impatto. L'eco di questo risultato si propagò rapidamente alle scienze biomediche, all'economia sperimentale, alla neuroimmagine cognitiva e ad altri campi empirici, innescando un dibattito metodologico di ampiezza inusuale. Ma la particolare forma che la crisi assume nell'intelligenza artificiale e nell'apprendimento automatico — il dominio che qui interessa da una prospettiva non soltanto tecnica ma filosofica — ha caratteri propri che la rendono distinta dalla crisi nelle scienze sperimentali tradizionali, più profonda per alcuni aspetti e più oscura per altri.

Nelle scienze sperimentali tradizionali, la crisi è prevalentemente riconducibile a pratiche statistiche scorrette: il **p-hacking** — la ricerca iterativa di configurazioni di variabili, sottocampioni o test statistici che producano un p-value al di sotto della soglia convenzionale di 0.05 — e il **publication bias** — la tendenza sistematica a pubblicare risultati positivi e a non pubblicare i negativi — producono una letteratura in cui la frequenza dei risultati significativi supera drammaticamente quella attesa sotto la sola ipotesi nulla. La stima di John Ioannidis (2005), nell'articolo «Why Most Published Research Findings Are False», mostrò mediante un argomento bayesiano semplice ma devastante che, in condizioni tipiche per molti campi biomedici (bassa prevalenza dell'ipotesi vera nella fase di esplorazione, piccole dimensioni campionarie, bias di pubblicazione), la **probabilità posteriore** che un risultato pubblicato come significativo sia effettivamente vero è inferiore al 50%. Il modello formale è il seguente: detto R il rapporto tra ipotesi vere e false nel campo prima dei test, α la probabilità di falso positivo, β la probabilità di falso negativo, u il coefficiente di bias, e n il numero di team che testano la stessa ipotesi, la probabilità che un risultato positivo pubblicato sia vero è:

$$PPV = (1 - \beta)R \,/\, [(1 - \beta)R + \alpha]$$

che, per valori tipici come $R = 0.1$ (una vera ogni dieci false), $\alpha = 0.05$, $\beta = 0.20$, dà PPV ≈ 0.44: meno della metà dei risultati pubblicati corrisponde a un effetto reale. L'aggiunta del bias u (la proporzione di studi distorti nel senso di segnalare risultati positivi indipendentemente dalla verità) abbassa ulteriormente PPV:

$$\text{PPV_bias} = [(1 - \beta)R + u\beta R] \,/\, [(1 - \beta)R + u\beta R + \alpha + u(1-\alpha)]$$

per $u = 0.3$ e gli stessi valori: $\text{PPV_bias} \approx 0.29$.

Nell'apprendimento automatico, i meccanismi della crisi sono parzialmente diversi ma concorrenti nella produzione degli stessi effetti. L'equivalente del p-hacking è il fenomeno del **benchmark overfitting**: la pratica, spesso inconsapevole ma strutturalmente incentivata dal sistema di pubblicazione, di ottimizzare le scelte architetturali, le procedure di addestramento, gli iperparametri e i criteri di valutazione sul benchmark di riferimento del campo fino al punto in cui le prestazioni sul benchmark migliorano senza che migliorino le capacità generalizzate del sistema. Il ciclo si autorinforza: i benchmark di riferimento — ImageNet per il riconoscimento di immagini, GLUE e SuperGLUE per la comprensione del linguaggio naturale, SQuAD per la question answering, MMLU per le conoscenze enciclopediche — diventano gli obiettivi di ottimizzazione impliciti dell'intera comunità, che pubblica progressi su di essi senza che questo comporti progressi equivalenti nelle applicazioni reali.

Dataset shift, distribuzione interna ed esterna, e il problema della generalizzazione

La nozione tecnica che incardina la maggior parte delle difficoltà epistemiche dell'apprendimento automatico è quella di **distribuzione dei dati**. Un sistema di apprendimento è addestrato su un insieme finito di esempi (il training set) estratti da una distribuzione P_{train} e valutato su un insieme di esempi (il test set) estratti da una distribuzione P_{test}. L'assunzione fondamentale del paradigma dell'apprendimento statistico — la **i.i.d. assumption** — è che i dati di training e di test siano **indipendenti e identicamente distribuiti**: $P_{train} = P_{test}$. Sotto questa assunzione, la teoria della generalizzazione — dai classici risultati di Vapnik e Chervonenkis sul legame tra il numero di campioni, la complessità del modello e il rischio empirico e atteso, ai risultati più recenti sulla generalizzazione dei modelli sovraparametrizzati — fornisce garanzie formali sull'errore che ci si attende su dati nuovi della stessa distribuzione.

In termini formali, il rischio atteso (errore reale) di un classificatore h su una distribuzione P è:

$$R(h) = \mathbb{E}_{(x,y)\sim P}[\ell(h(x), y)]$$

mentre il rischio empirico su un campione $S = \{(x_1,y_1), \ldots, (x_n,y_n)\}$ di n esempi i.i.d. da P è:

$$\hat{R}_S(h) = (1/n) \sum_i \ell(h(x_i), y_i)$$

Il teorema fondamentale dell'apprendimento statistico nel PAC learning (Probably Approximately Correct, Valiant 1984) afferma che, per una classe di ipotesi H di dimensione VC finita d_VC, con probabilità almeno 1-δ su un campione di n esempi i.i.d.:

$$R(h) \leq \hat{R}_S(h) + \sqrt{\left([d_{VC} (\log(2n/d_{VC}) + 1) + \log(4/\delta)] / n \right)}$$

Il bound di generalizzazione fornisce una garanzia sull'errore atteso in termini del rischio empirico, della dimensione VC e del numero di campioni. Ma il bound è valido soltanto sotto la i.i.d. assumption: se i dati di test hanno una distribuzione diversa da quelli di training, nessuna garanzia del genere si applica.

Il **dataset shift** — la violazione dell'assunzione i.i.d. — assume diverse forme che la letteratura ha classificato e analizzato. Il **covariate shift** è il caso in cui la distribuzione degli input cambia ($P_{test}(x) \neq P_{train}(x)$) ma la distribuzione condizionale delle etichette dato l'input rimane la stessa ($P_{test}(y|x) = P_{train}(y|x)$). Tecniche di **importance weighting** — che pesano i dati di training in proporzione al rapporto $P_{test}(x)/P_{train}(x)$ — possono correggere il covariate shift se le due distribuzioni hanno supporto sovrapponente; ma stimare i pesi richiede conoscenza o stima di P_{test}, che spesso non è disponibile. Il **concept drift** è il caso in cui cambia la distribuzione condizionale $P(y|x)$ nel tempo — il significato dei segnali cambia, come avviene nei mercati finanziari o nei trend del linguaggio naturale. Il **prior probability shift** è il caso in cui cambia la distribuzione delle etichette $P(y)$ mantenendo costante $P(x|y)$: rilevante in diagnostica medica dove la prevalenza delle malattie varia tra popolazioni diverse.

La distinzione epistemicamente più profonda — e quella che ha maggiori implicazioni per la valutazione dei sistemi di apprendimento automatico — è quella tra **generalizzazione intra-distribuzionale** e **generalizzazione extra-distribuzionale** (o **out-of-distribution generalization**, OOD). La prima è la capacità di fare previsioni corrette su esempi nuovi della stessa distribuzione dei dati di training: è ciò che la teoria PAC e le sue estensioni misurano, ed è il tipo di generalizzazione che i benchmark standard valutano. La seconda è la capacità di fare previsioni corrette su esempi che provengono da distribuzioni diverse da quella di training: è la capacità che si richiede in applicazioni reali, dove la distribuzione del deployment raramente coincide con quella dell'addestramento.

I progressi spettacolari dei modelli di apprendimento automatico sugli ultimi quindici anni — la riduzione del tasso di errore sul benchmark ImageNet dal 26% nel 2011 (anno di introduzione del dataset) all'1.8% nel 2021 (Zhai et al., ViT-G/14) — riguardano quasi esclusivamente la generalizzazione intra-distribuzionale. Quando si testano gli stessi modelli su varianti di ImageNet costruite per mettere alla prova la generalizzazione OOD — ImageNet-C (corruptions, Hendrycks & Dietterich, 2019: trasformazioni come rumore gaussiano, sfocatura, variazioni di contrasto), ImageNet-A (campioni difficili selezionati per confondere classificatori addestrati su ImageNet), ImageNet-Sketch (versioni stilizzate dei soggetti) — le prestazioni degradano in modo drammatico: i modelli che hanno un errore inferiore al 2% su ImageNet standard mostrano errori del 30–60% su ImageNet-C e ImageNet-A. Questo

divario — grande, sistematico e non colmato dall'aumento della scala — è la forma concreta che la crisi della generalizzazione assume nell'apprendimento automatico.

Il benchmark overfitting come problema epistemico: validità ecologica e cicli di retroazione

Il meccanismo del benchmark overfitting merita un'analisi filosofica più approfondita di quanto la letteratura tecnica di solito gli dedichi, perché esso è la manifestazione nell'apprendimento automatico di una difficoltà epistemica generale che riguarda la relazione tra misura e realtà misurata. Il processo attraverso cui un benchmark di valutazione perde progressivamente il proprio valore informativo è formalmente analogo al fenomeno descritto dalla **Legge di Goodhart**: «quando una misura diventa un obiettivo, cessa di essere una buona misura». Formulata originariamente nell'economia monetaria da Charles Goodhart (1975) a proposito degli indicatori macroeconomici usati dalle banche centrali come obiettivi di policy, la legge cattura il fatto che qualsiasi misura è valida come surrogato di un costrutto di interesse soltanto nella misura in cui non è oggetto di ottimizzazione diretta; l'ottimizzazione del surrogato invece del costrutto produce un sistema che massimizza la misura senza migliorare il costrutto.

Nel caso dei benchmark dell'apprendimento automatico, il costrutto di interesse è la capacità generalizzata del sistema su distribuzioni di dati reali; la misura surrogata è l'accuratezza su un dataset fisso. Nel momento in cui il benchmark diventa il metro di paragone istituzionalizzato per la comparazione dei metodi — il criterio di accettazione degli articoli, il riferimento nelle proposte di finanziamento, il target dei sistemi di Neural Architecture Search — la pressione di ottimizzazione si concentra su di esso, e i progressi sul benchmark divergono dai progressi sulla capacità generalizzata. La comunità scientifica adatta i propri metodi al benchmark, non al problema; il benchmark diventa una nicchia ecologica a cui i metodi si adattano, non un ambiente rappresentativo del problema reale.

Il termine tecnico per la violazione della corrispondenza tra la misura e il costrutto d'interesse è **validità ecologica** — concetto elaborato nella psicologia sperimentale da Egon Brunswik negli anni Quaranta per criticare i paradigmi di laboratorio che ottimizzano il controllo sperimentale a scapito della rappresentatività ecologica. Nella psicologia, un esperimento ha alta validità ecologica se le condizioni del laboratorio sono rappresentative delle condizioni naturali in cui avviene il comportamento di interesse; ha bassa validità ecologica se le condizioni di laboratorio sono così artificiali che i risultati non si generalizzano fuori dal laboratorio. Il trasferimento diretto di questo concetto all'apprendimento automatico dà: un benchmark ha alta validità ecologica se è rappresentativo della distribuzione su cui il sistema sarà effettivamente deployato; ha bassa validità ecologica se è costruito in modo da facilitare la misurazione e la comparazione a scapito della rappresentatività.

La misura esatta del divario di validità ecologica di un benchmark è difficile da stabilire a priori, ma esistono metodologie post-hoc per stimarla. Benjamin Recht e collaboratori (2019) hanno costruito nuovi test set per ImageNet e CIFAR-10 — ImageNet-v2 e CIFAR-10.1 — seguendo scrupolosamente la procedura originale di costruzione dei dataset ma raccogliendo nuovi campioni: i modelli addestrati su ImageNet mostravano un calo sistematico dell'accuratezza di 3–15 punti percentuali su ImageNet-v2 rispetto a ImageNet standard, anche senza nessun cambiamento nella distribuzione nominale del task. Il calo è attribuibile a una combinazione di sottile distributional shift nel processo di raccolta e di overfitting sull'idiosincrasia del dataset originale — le decisioni di annotazione, le fonti delle immagini, le convenzioni fotografiche — che i modelli avevano appreso senza che questo fosse intenzionale o consapevole per i ricercatori.

La questione degli iperparametri e la riproducibilità computazionale

Un livello più basso della crisi della riproducibilità nell'apprendimento automatico riguarda la **riproducibilità computazionale** in senso stretto: la capacità di ottenere gli stessi risultati numerici a partire dallo stesso codice, sugli stessi dati, con le stesse specifiche hardware. Questa forma di riproducibilità — distinta dalla riproducibilità scientifica, che richiede che i risultati si mantengano su dati nuovi con metodi analoghi — è in linea di principio risolvibile attraverso la fissazione dei seed random, la versioning del codice e delle dipendenze, la documentazione delle configurazioni hardware. In pratica, è sistematicamente violata.

Una fonte strutturale di irriproducibilità computazionale è il non-determinismo delle implementazioni di operazioni parallele su GPU. Molte operazioni di algebra lineare implementate nelle librerie CUDA (per le GPU NVIDIA) e nelle equivalenti per altre architetture sono non-deterministiche per design: l'ordine in cui le operazioni in virgola mobile vengono eseguite in parallelo varia da un'esecuzione all'altra in modo dipendente dal scheduling dello hardware, e poiché l'aritmetica in virgola mobile non è associativa — $(a + b) + c \neq a + (b + c)$ in generale per numeri a precisione finita — il risultato varia tra esecuzioni diverse. L'entità delle variazioni è tipicamente piccola per ogni singola operazione, ma si accumula su milioni di operazioni durante il training e può produrre, per determinati regimi di addestramento (tassi di apprendimento elevati, loss landscapes mal condizionati), differenze macroscopiche nelle reti risultanti.

Il **problema degli iperparametri** è la dimensione della crisi di riproducibilità più direttamente rilevante per la comparazione dei metodi. Un sistema di apprendimento automatico dipende da un numero tipicamente elevato di scelte di design: l'architettura della rete, il tasso di apprendimento e la sua schedulazione, l'ottimizzatore e i suoi parametri (momentum, weight decay, coefficienti Adam β_1 e β_2), le procedure di data augmentation, le dimensioni dei batch, le strategie di regolarizzazione (dropout, batch normalization, label smoothing), le inizializzazioni dei pesi. Per ogni combinazione di queste scelte, il sistema addestrato ha prestazioni diverse; il spazio degli iperparametri è tipicamente

continuo e di alta dimensionalità; la funzione che mappa gli iperparametri sulle prestazioni finali non è convessa né facilmente caratterizzabile analiticamente.

La pratica standard di reporting dei risultati consiste nel presentare le prestazioni del modello con gli iperparametri migliori trovati attraverso una ricerca estensiva (grid search, random search, ottimizzazione bayesiana degli iperparametri). Ma se la ricerca degli iperparametri non è riportata e replicata, il risultato dipende dall'intensità della ricerca: un metodo con iperparametri subottimali e un secondo metodo con iperparametri ottimali non sono comparabili, perché il divario di prestazioni potrebbe riflettere la differenza nella ricerca degli iperparametri invece che la differenza intrinseca tra i metodi. Dodge e colleghi (2019) e Henderson e colleghi (2018) hanno dimostrato empiricamente, su task di NLP e di reinforcement learning rispettivamente, che variazioni negli iperparametri producono variazioni di prestazione dell'ordine di grandezza dei progressi riportati come avanzamenti dello stato dell'arte: metodi presentati come superiori si rivelano inferiori o comparabili una volta che la ricerca degli iperparametri è equiparata.

VI.6.5 — Selezione del modello, test multipli e l'inflazione dell'errore di tipo I

La questione statistica alla base di molte forme di irriproducibilità nell'apprendimento automatico è quella della **selezione del modello** e del controllo dell'errore in presenza di test multipli. Se si confrontano k metodi su m benchmark, si eseguono implicitamente k×m confronti statistici; se si selezionano per la pubblicazione soltanto i confronti con risultati favorevoli al metodo proposto, si esegue implicitamente un'esplorazione massiva dello spazio dei confronti e si riporta soltanto l'estremo favorevole. Il tasso di errore di tipo I — la probabilità di osservare una differenza significativa per caso sotto l'ipotesi nulla che i metodi siano equivalenti — si gonfia in proporzione al numero di confronti: se $\alpha = 0.05$ per ogni confronto, la probabilità di almeno un falso positivo su m confronti indipendenti è

$$1 - (1 - \alpha)^m$$

che per $m = 50$ è circa 0.92: quasi certamente si osserverà almeno un confronto significativo per caso. La correzione di Bonferroni — sostituire α con α/m — corregge il family-wise error rate ma è conservativa, in quanto assume indipendenza tra i test. La correzione di Benjamini-Hochberg controlla invece il **False Discovery Rate** (FDR) — la proporzione attesa di falsi positivi tra tutti i positivi riportati — a un livello q specificato: ordina i p-value $p_{(1)} \le p_{(2)} \le \dots \le p_{(m)}$ e definisce $k^* = \max\{i : p_{(i)} \le iq/m\}$; si rigettano le prime k^* ipotesi nulle. Questa procedura garantisce FDR $\le q$ con minore perdita di potenza rispetto a Bonferroni.

Ma queste correzioni sono raramente applicate nella letteratura dell'apprendimento automatico, in parte perché i confronti non sono esplicitati come test statistici multipli ma come sperimentazione esplorativa, e in parte perché la cultura del campo — orientata più all'ingegneria che alla statistica

inferenziale — non ha internalizzato le pratiche del testing multiplo come standard metodologico. Il risultato è un corpus di letteratura in cui i confronti riportati sono sistematicamente selezionati e i risultati superiori allo stato dell'arte sono sistematicamente sovrastimati.

Il problema della validità costruttiva nell'IA: cosa misurano i benchmark?

La dimensione più profonda e meno tecnica della crisi della riproducibilità nell'IA è quella che nella psicologia psicometrica si chiama **validità costruttiva** (*construct validity*): la questione di se uno strumento di misurazione misura effettivamente il costrutto che intende misurare, o misura qualcosa d'altro correlato al costrutto. La validità costruttiva di un benchmark richiede che i task che lo compongono siano effettivamente rappresentativi del costrutto di interesse — per esempio, la «comprensione del linguaggio naturale» (NLU) — e che le prestazioni su di essi riflettano genuinamente le capacità cognitive che il termine nomina.

I lavori di Gururangan e colleghi (2018) e di Poliak e colleghi (2018) hanno dimostrato che modelli addestrati soltanto sulle ipotesi di dataset di NLI (*Natural Language Inference*) — senza accesso alle premesse rispetto a cui le ipotesi devono essere valutate — raggiungono prestazioni superiori al 65% su alcuni dataset standard, significativamente sopra la baseline casuale del 33%. Questo significa che le ipotesi del dataset contengono **artefatti statistici** — correlazioni spurie tra le caratteristiche superficiali delle ipotesi e le loro etichette — che un classificatore può sfruttare senza svolgere alcun ragionamento inferenziale. Le prestazioni elevate dei sistemi NLI non misurano la capacità di inferire relazioni logiche tra premesse e ipotesi: misurano la capacità di sfruttare artefatti del dataset.

Analogamente, McCoy e colleghi (2019) hanno mostrato che modelli BERT fine-tuned su SNLI e MultiNLI — dataset di NLI di riferimento — mostrano performance degradate su HANS (*Heuristic Analysis for NLI Systems*), un dataset di diagnostica costruito per testare tre euristiche superficiali (sovrapposizione lessicale, sottosequenza, struttura costituente) che non sono valide inferenzialmente ma che predicono l'etichetta nel dataset standard con alta accuratezza. I modelli che usano le euristiche ottengono alte prestazioni su SNLI ma bassissime su HANS; i modelli che non le usano sono più lenti da addestrare ma più robusti. Il dataset standard, in questo senso, seleziona contro la robustezza.

Per i modelli linguistici di larga scala come GPT-4 o i sistemi della serie Claude, la questione della validità costruttiva diventa ancora più acuta e più filosoficamente interessante. I benchmark come MMLU (*Massive Multitask Language Understanding*) — che testa le prestazioni su 57 task che coprono discipline accademiche dalla medicina alla fisica al diritto — sono stati costruiti prima dell'era dei modelli di larga scala come strumenti di valutazione del progresso verso la comprensione generalizzata. Ma i modelli addestrati su enormi corpora testuali hanno quasi certamente visto, durante il training, testi che descrivono o che contengono le domande dei benchmark o formulazioni molto simili a esse: il **data**

contamination — la presenza nel corpus di addestramento di versioni delle domande di test — rende impossibile distinguere tra comprensione genuina e memorizzazione diretta o per analogia superficiale.

La difficoltà di stabilire la contaminazione del training corpus è strutturale: i corpora di addestramento dei modelli commerciali non sono pubblicamente disponibili nella loro interezza, e la verifica dell'assenza di contaminazione richiederebbe l'accesso completo ai dati e ai metadati di preprocessing. Golchin e Surdeanu (2023) hanno sviluppato metodi per rilevare la contaminazione attraverso la probabilità assegnata dal modello a sequenze di dati di test versus sequenze permutate; Magar e Schwartz (2022) hanno mostrato che i modelli tendono a mostrare prestazioni superiori sui dataset con cui sembrano familiari rispetto alle varianti con le stesse proprietà strutturali ma distribuzione diversa. Ma questi test indiretti non sono definitivi, e la questione della contaminazione rimane una fonte sistematica di incertezza sulla validità dei benchmark per i modelli di larga scala.

Implicazioni epistemiche: il problema della generalizzazione extra-distribuzionale come questione filosofica

La crisi della riproducibilità nell'IA non è soltanto una questione di pratiche di ricerca scorrette che si possono correggere con metodologie migliori: è la manifestazione di una difficoltà epistémica di fondo che riguarda la natura della **conoscenza in un sistema appreso statisticamente**.

Un sistema di apprendimento automatico apprende una funzione dal training set. La funzione appresa è, nel migliore dei casi, una buona approssimazione della funzione «vera» sui punti della distribuzione di training; ma non esiste — in linea di principio — alcuna ragione per cui debba generalizzare a distribuzioni diverse, a meno che si supponga che la funzione «vera» abbia proprietà strutturali (regolarità, semplicità, invarianza a certe trasformazioni) che si riflettono nella funzione appresa. Il bias induttivo del sistema — l'insieme di assunzioni implicite sulla struttura della funzione da apprendere che sono codificate nell'architettura, nelle procedure di training e nelle funzioni di loss — determina quali distribuzioni diverse da quella di training il sistema è capace di trattare con accuratezza. Nella misura in cui il bias induttivo corrisponde alla struttura reale del problema — come fa la convoluzione per le immagini naturali, sfruttando l'invarianza traslazionale — il sistema generalizza bene OOD; nella misura in cui non corrisponde, il sistema fallisce.

Ma il bias induttivo non è mai esplicito e raramente esaminato criticamente. I modelli linguistici di larga scala hanno bias induttivi enormemente complessi, determinati dalla combinazione dell'architettura transformer, delle procedure di training (pretraining, RLHF, constitutional AI), delle dimensioni del modello e della scala del corpus. Questi bias determinano quali trasformazioni del dominio di input il modello tratta come equivalenti (invarianza), quali strutture il modello tende a privilegiare nelle sue previsioni (prior implicito), e in quale senso le previsioni del modello possono essere considerate appropriate su distribuzioni diverse da quella di training. Nessuno di questi aspetti è formalmente specificato

o formalmente verificabile; la loro caratterizzazione empirica attraverso test OOD è necessaria ma non sufficiente, perché ogni test OOD testa una specifica deviazione dalla distribuzione di training e non può coprire l'insieme potenzialmente infinito delle deviazioni possibili.

Il problema acquista dimensione filosofica nella sua connessione con il **problema dell'induzione** nel senso di Hume. Hume aveva mostrato che nessun numero finito di osservazioni può fornire una giustificazione logicamente conclusiva per una legge universale: l'inferenza induttiva non è giustificata dalla logica ma dalla natura e dall'abitudine. Nella versione che emerge dalla crisi della riproducibilità nell'IA, il problema si precisa: non soltanto non c'è giustificazione logica per generalizzare da un training set finito a una distribuzione futura; non c'è nemmeno garanzia empirica che la generalizzazione avvenga in modo utile al di fuori della distribuzione di addestramento. I sistemi di apprendimento automatico sono, nel senso più preciso, macchine induttive: le loro garanzie di generalizzazione sono valide soltanto nella misura in cui il futuro assomiglia al passato — nella misura in cui P_deploy = P_train. Quando questa assunzione è violata — come accade quasi universalmente nelle applicazioni reali, dove le distribuzioni cambiano nel tempo, nello spazio e tra contesti — le garanzie si dissolvono.

La risposta tecnica a questa situazione è il programma della **robust machine learning** e del **domain generalization**: la costruzione di sistemi che generalizzano non ad una distribuzione specifica ma a famiglie di distribuzioni, sfruttando invarianze strutturali del problema. Il metodo IRM (*Invariant Risk Minimization*, Arjovsky et al., 2019) formula il problema come la ricerca di una rappresentazione $\varphi: X \to Z$ e di un classificatore $w: Z \to Y$ tali che il classificatore w sia simultaneamente ottimale su tutte le distribuzioni di training disponibili:

$$\min_{\varphi, w} \sum_e R^e(w \circ \varphi) \quad \text{soggetto a: } w \in \operatorname{argmin}_{\tilde{w}} R^e(\tilde{w} \circ \varphi) \text{ per ogni ambiente } e$$

dove R^e è il rischio sull'ambiente (distribuzione) e. Il razionale è che, se w è ottimale su tutte le distribuzioni osservate, la relazione che implementa è probabilmente invariante rispetto alle distribuzioni non osservate — è probabile che stia sfruttando correlazioni stabili invece di correlazioni spurie specifiche di alcuni ambienti. Il metodo IRM e i suoi successori (IRM-v2, DIGA, VREx) producono miglioramenti empirici sulla generalizzazione OOD in alcuni domini; ma i miglioramenti non sono né sistematici né privi di eccezioni, e il problema teorico di come identificare le invarianze rilevanti senza conoscere le distribuzioni di test rimane aperto.

La crisi della riproducibilità nell'IA è dunque, nella sua dimensione più profonda, la forma contemporanea di un problema antico: il problema di come estrarre conoscenza universale da osservazioni particolari, di come distinguere la struttura reale dal rumore contingente del campione, di come costruire sistemi che conoscono il mondo invece di sistemi che memorizzano le proprie esperienze. La differenza rispetto all'epistemologia tradizionale è che i sistemi di apprendimento automatico rendono questo problema computazionalmente concreto e massivamente scalabile: le domande che Hume e Popper ponevano in astratto sul

ragionamento induttivo si presentano ora come problemi tecnici urgenti con implicazioni dirette per la sicurezza e l'affidabilità di sistemi che influenzano decisioni in ambito medico, giuridico, finanziario e militare. L'epistemologia dell'IA non è, sotto questo profilo, un esercizio accademico: è la riflessione sulle condizioni di possibilità della conoscenza in sistemi che la società ha già ampiamente deployato.

APPENDICE III: Semantica Vettoriale, Word Embeddings e la Geometria dello Spazio Semantico Distribuzionale

Il punto di congiunzione

Vi è un punto nella storia recente della linguistica computazionale in cui una tradizione filosofica pluridecennale — quella della semantica distribuzionale fondata sull'ipotesi harrissiana e sul secondo Wittgenstein — ha incontrato una tradizione matematica di lungo corso — quella dell'algebra lineare applicata alle statistiche del linguaggio — producendo una classe di modelli che costituisce, per più aspetti, il punto di contatto più diretto tra la filosofia del significato e l'architettura dei sistemi artificiali moderni. Questo punto di incontro è la semantica vettoriale degli word embeddings: la famiglia di rappresentazioni che include Word2Vec, GloVe, fastText e le loro estensioni contestuali, e che costituisce il substrato matematico su cui i transformer moderni — attraverso la meccanica dell'attenzione e del residual stream esaminata nella sezione IV.3 — costruiscono rappresentazioni di ordine superiore.

La sezione II.7 ha introdotto la semantica distribuzionale nel suo quadro formale di base: le matrici di co-occorrenza, la PMI, la riduzione attraverso SVD, la similarità coseno. Questa sezione si propone di approfondire l'analisi su tre piani connessi ma distinti. Il primo è filosofico: l'esame accurato di come la tradizione inaugurata da Zellig Harris e la filosofia tardiva di Wittgenstein convergano sull'ipotesi distribuzionale e di cosa questa convergenza implichi sul piano dell'ontologia del significato. Il secondo è matematico: la derivazione dettagliata degli obiettivi di addestramento e delle proprietà geometriche di Word2Vec e GloVe, con attenzione alle connessioni teoriche tra i due modelli e alla struttura dello spazio semantico che generano. Il terzo è filosofico-empirico: l'analisi delle proprietà strutturali dello spazio semantico distribuzionale — le analogie vettoriali, le direzioni semantiche lineari, la geometria delle relazioni categoriali — e di cosa queste proprietà implichino per la comprensione filosofica del significato.

Harris e il distribuzionalismo: la struttura formale dell'ipotesi

Zellig Harris è una figura di straordinaria importanza nella storia della linguistica del Novecento, oscurata dall'enorme ombra proiettata dal suo allievo Noam

Chomsky. Mentre Chomsky sviluppava il programma generativista basato sulla competenza sintattica innata, Harris elaborava un programma radicalmente diverso: l'analisi distribuzionale del linguaggio come metodo per la descrizione strutturale delle lingue naturali senza fare appello a significati intesi come entità extralinguistiche.

Il saggio fondativo è «Distributional Structure» del 1954, pubblicato su *Word*. Il nucleo dell'ipotesi è enunciato in termini che meritano una citazione accurata della struttura argomentativa, anche se non del testo letterale: le parole che compaiono nelle stesse posizioni nell'ambiente linguistico — negli stessi contesti di distribuzione — tendono ad avere significati simili o equivalenti; viceversa, le parole semanticamente simili tendono a comparire negli stessi contesti. La similarità di distribuzione è una proxy della similarità semantica: non perché il significato sia identico alla distribuzione, ma perché la distribuzione è la traccia linguistica del significato.

Il programma harrissiano è metodologicamente più cauteloso di quanto l'ipotesi distribuzionale possa sembrare. Harris non afferma che il significato sia riducibile alla distribuzione; afferma che la distribuzione è l'unico dato osservabile direttamente disponibile all'analista linguistico, e che la struttura semantica può essere recuperata dalla struttura distribuzionale. Questo è un programma epistemologico — sulla conoscibilità del significato a partire dai dati linguistici — non necessariamente un programma ontologico sulla natura del significato.

La formalizzazione dell'ipotesi distribuzionale richiede la definizione precisa di «distribuzione» e di «contesto». Siano dati un corpus $C = (w_1, w_2, ..., w_N)$ di N token e un vocabolario $V = \{w_1, ..., w_\{|V|\}\}$. Il **contesto sintagmatico** di occorrenza di un token w_t è la finestra di m token che lo circonda: $(w_\{t-m\}, ..., w_\{t-1\}, w_\{t+1\}, ..., w_\{t+m\})$. La **distribuzione contestuale** di una parola $w \in V$ è la distribuzione di probabilità $P(c \mid w)$ sui contesti c, stimata dal corpus:

$$P(c \mid w) \approx f(w, c) / f(w)$$

dove $f(w, c)$ è la frequenza di co-occorrenza di w con il contesto c, e $f(w) = \Sigma_c f(w, c)$ è la frequenza marginale di w. La **similarità distribuzionale** tra due parole w_1 e w_2 è allora una misura della prossimità delle loro distribuzioni contestuali, tipicamente la distanza tra i vettori che le rappresentano.

L'ipotesi harrissiana afferma che:

$$\text{sim_semantica}(w_1, w_2) \approx \text{sim_distribuzionale}(w_1, w_2) = S(P(c|w_1), P(c|w_2))$$

per qualche misura di similarità S tra distribuzioni. Questa è una tesi empirica, non analitica: la sua verità o falsità dipende dal grado in cui le strutture distribuzionali del linguaggio riflettono le strutture semantiche. Le evidenze accumulate nel corso di decenni di ricerca in linguistica computazionale sono schiaccianti nel sostenere l'ipotesi per vaste classi di relazioni semantiche: la similarità lessicale (sinonimi, antonimi, iponimi), le relazioni categoriali (appartenenza a categorie semantiche), le relazioni tematiche (co-partecipazione a scenari o frame semantici), le relazioni analogiche (coppie che esibiscono relazioni proporzionali del tipo «maschio:femmina = re:regina»).

Harris ha inoltre sviluppato, negli anni Sessanta e Settanta, un'analisi più raffinata in termini di **trasformazioni linguistiche** e di **operatori**: le trasformazioni che preservano il significato — le relazioni parafrastiche tra frasi — corrispondono a operazioni regolari sulla struttura distribuzionale. Una parola è definita non soltanto dalla distribuzione dei suoi contesti ma dal sistema di relazioni trasformazionali in cui figura. Questo approfondimento apre una direzione verso una semantica distribuzionale relazionale che anticipa, in forma qualitativa, i meccanismi di attenzione dei transformer.

Wittgenstein e il significato come uso: la radice filosofica dell'ipotesi distribuzionale

La connessione tra l'ipotesi distribuzionale di Harris e la filosofia tardiva di Ludwig Wittgenstein non è soltanto una somiglianza superficiale tra posizioni che condividono uno slogan — «il significato è l'uso», «il significato è determinato dalla distribuzione dei contesti» — ma una convergenza profonda su concezioni radicalmente antiessenzialiste del significato che merita un'analisi filosofica mente attenta.

Le *Philosophische Untersuchungen* (pubblicate postume nel 1953, l'anno precedente a «Distributional Structure») costituiscono il documento più elaborato di una delle più radicali rivoluzioni nella storia della filosofia del linguaggio: il rifiuto della concezione **agostiniana** del significato — quella secondo cui ogni parola ha un significato che è l'oggetto che la parola designa — e la sua sostituzione con una concezione del significato come **uso** regolato all'interno di pratiche linguistiche condivise.

Il §43 delle *Ricerche* enuncia il principio che è diventato il locus classico della filosofia del linguaggio wittgensteiniana: «Il significato di una parola è il suo uso nel linguaggio» (*Die Bedeutung eines Wortes ist sein Gebrauch in der Sprache*). Questa proposizione, nella lettura corretta, non è una definizione ma una remissione: quando ci si chiede cosa sia il significato di una parola, si deve guardare a come la parola è effettivamente usata nelle pratiche linguistiche, non cercare un'entità mentale o platonica cui essa corrisponde. Il significato non è un'immagine nella mente, non è un oggetto nel mondo, non è una condizione di verità astratta: è la regola dell'uso della parola nei giochi linguistici in cui essa compare.

La nozione di **gioco linguistico** (*Sprachspiel*) è la categoria strutturale fondamentale della semantica wittgensteiniana. Un gioco linguistico è una pratica linguistica intrecciata con attività non linguistiche — forme di vita — all'interno della quale le espressioni linguistiche acquistano i loro ruoli funzionali. «Lasciami passare!» nel gioco degli scacchi ha un significato diverso da «Lasciami passare!» in una disputa di precedenza stradale: non perché le parole siano diverse, ma perché i giochi linguistici in cui esse compaiono sono diversi, con regole diverse, scopi diversi, conseguenze diverse.

La connessione con l'ipotesi distribuzionale è ora visibile: il «contesto di occorrenza» harrissiano è una versione formalizzata e matematicamente operazionalizzabile del «gioco linguistico» wittgensteiniano. Harris formalizza la

distribuzione come insieme di co-occorrenze statistiche; Wittgenstein la concettualizza come insieme di pratiche regolate. In entrambi i casi, il significato è una proprietà relazionale dell'espressione rispetto al sistema di usi o di contesti in cui compare, non una proprietà intrinseca dell'espressione stessa.

La nozione wittgensteiniana di **rassomiglianze di famiglia** (*Familienähnlichkeiten*, §§65-67) è filosofica mente connessa alla geometria degli spazi di embedding. Wittgenstein osserva che molte categorie semantiche non hanno una proprietà comune che definisce l'appartenenza alla categoria: i «giochi» — i giochi da tavolo, i giochi di carte, i giochi sportivi, i giochi di parole — non condividono una proprietà essenziale. Sono tenuti insieme da una rete di rassomiglianze sovrapposte e incrociate: «come i diversi membri di una famiglia si rassomigliano per la costituzione fisica, la forma del viso, il colore degli occhi, il portamento, il temperamento». Non vi è una proprietà che tutti i giochi hanno in comune; vi è una struttura di sovrapposizioni di rassomiglianze che costituisce la categoria.

La geometria degli spazi semantici distribuzionali realizza, in modo formale, questa struttura di rassomiglianze di famiglia. La categoria dei «giochi» non corrisponde, nello spazio degli embedding, a una regione delimitata da una frontiera netta e da un centro prototipico unico; corrisponde a una regione di forma irregolare, con sottocluster che si sovrappongono parzialmente — giochi da tavolo vicini ai giochi di carte, giochi di carte vicini ai giochi di parole, giochi di parole lontani dai giochi da tavolo — la cui coesione è data dalla densità locale delle rassomiglianze piuttosto che dalla condivisione di proprietà essenziali.

L'analisi wittgensteiniana della **seguire una regola** (*Regelfolgen*, §§138-242) ha implicazioni più radicali per la semantica distribuzionale. Wittgenstein osserva che qualsiasi comportamento passato è compatibile con infinite regole diverse: nessuna sequenza finita di esempi determina univocamente la regola che li genera, perché infiniti modi di continuare la sequenza sono tutti «compatibili con la regola» nel senso che si può trovare una formulazione della regola che li giustifichi. Questo è il nucleo del problema scettico sollevato da Kripke nella sua lettura di Wittgenstein (*Wittgenstein on Rules and Private Language*, 1982): cosa costituisce il fatto che una persona intenda una regola piuttosto che un'altra, dato che il comportamento passato sottodetermina la regola?

Questo problema ha un'eco diretta nel problema del grounding degli embedding distribuzionali: cosa fa sì che il vettore di embedding della parola «verde» si riferisca al verde piuttosto che al grue (la proprietà di essere verde prima di una certa data e blu dopo)? Il vettore è determinato dalla distribuzione delle co-occorrenze nel corpus, e questa distribuzione è compatibile con infinite ipotesi interpretative diverse. Il sistema di embedding non contiene, al suo interno, un fatto che determini quale delle interpretazioni compatibili è quella corretta. La risposta harrissiana — che la distribuzione è sufficiente per fissare il significato — è vulnerabile alla stessa critica scettica che Kripke rivolge a Wittgenstein: la distribuzione sottodetermina la regola, e non vi è un fatto interno al sistema distribuzionale che selezioni l'interpretazione corretta.

La risposta wittgensteiniana al problema kripkeano — che la seguire una regola è una pratica comunitaria, non un fatto interno alla mente del singolo parlante —

ha una contropartita distribuzionale: il grounding degli embedding non è un fatto interno al sistema di embedding, ma è fornito dalla comunità di parlanti che ha prodotto il corpus. Il corpus non è soltanto un insieme di co-occorrenze statistiche; è il sedimento linguistico di una comunità di pratiche, e le pratiche hanno connessioni causali con il mondo che i puri dati statistici non catturano direttamente. Questa risposta è parzialmente convincente, ma non risolve il problema del grounding per i termini astratti e per le espressioni il cui uso è radicato in pratiche altamente specializzate.

La costruzione formale degli word embedding: Word2Vec

Il modello Word2Vec, introdotto da Mikolov, Chen, Corrado e Dean in «Efficient Estimation of Word Representations in Vector Space» (2013, arXiv:1301.3781) e nelle pubblicazioni successive, rappresenta un punto di discontinuità nella storia della semantica distribuzionale computazionale: per la prima volta, diventa possibile apprendere rappresentazioni vettoriali dense di alta qualità su corpora di miliardi di token in tempi computazionali praticabili su hardware comune. Il modello esiste in due varianti architetturali: **CBOW** (Continuous Bag of Words) e **Skip-gram**. L'analisi che segue si concentra su Skip-gram, che ha proprietà teoriche più eleganti e prestazioni empiriche superiori su rare words e relazioni analogiche.

Il modello Skip-gram associa a ogni parola $w \in V$ due vettori: un vettore di **embedding di input** $v_w \in \mathbb{R}^d$ (usato quando w compare come parola centrale) e un vettore di **embedding di output** $\tilde{u}_w \in \mathbb{R}^d$ (usato quando w compare come parola di contesto). L'obiettivo è la massimizzazione della log-verosimiglianza delle parole di contesto dato il termine centrale, calcolata sull'intero corpus:

$$J(\theta) = \frac{1}{T} \sum_{t=1}^{T} \sum_{j \in [-m,m],\, j \neq 0} \log P(w_{t+j} \mid w_t)$$

dove T è la lunghezza del corpus, m è la dimensione della finestra di contesto, e $\theta = \{v_w, \tilde{u}_w\}_{w \in V}$ è l'insieme dei parametri. La probabilità condizionale è modellata attraverso il softmax gerarchico o, nella variante più comunemente utilizzata, attraverso il softmax plano:

$$P(w_O \mid w_I) = \frac{\exp(\tilde{u}_{w_O}^{\top} \cdot v_{w_I})}{\sum_{w=1}^{|V|} \exp(\tilde{u}_w^{\top} \cdot v_{w_I})}$$

Il calcolo del denominatore di questa espressione — la normalizzazione sul vocabolario intero — richiede $O(|V|)$ operazioni per ogni coppia (w_I, w_O), il che è computazionalmente proibitivo per vocabolari di centinaia di migliaia di parole. Mikolov e colleghi propongono due approssimazioni efficienti.

La prima è il **softmax gerarchico** (*hierarchical softmax*): si costruisce un albero binario di Huffman sul vocabolario, in cui le parole frequenti hanno codici brevi e le parole rare hanno codici lunghi. La probabilità di una parola è espressa come prodotto di probabilità lungo il cammino dalla radice alla foglia corrispondente:

$$P(w_O \mid w_I) = \prod_{j=1}^{L(w_O)} \sigma([\![n(w_O, j+1) = \mathrm{ch}(n(w_O, j))]\!] \cdot \tilde{u}_{n(w_O,j)}^{\top} \cdot v_{w_I})$$

dove L(w_O) è la lunghezza del cammino verso w_O, n(w_O, j) è il j-esimo nodo lungo il cammino, ch(n) è il figlio sinistro di n, $[\![\cdot]\!]$ è la funzione indicatrice, e σ è la funzione sigmoide. La complessità di calcolo è O(log |V|) per coppia.

La seconda, e più popolare, è il **negative sampling**: invece di normalizzare sull'intero vocabolario, si campionano k parole «negative» w_1, ..., w_k dalla distribuzione rumorosa P_n(w) $\propto$ f(w)^{3/4} (dove f(w) è la frequenza di w nel corpus, elevata a 3/4 per ridurre il vantaggio delle parole molto frequenti), e si addestra il modello a distinguere la parola di contesto reale dai negativi campionati. L'obiettivo per ogni coppia osservata (w_I, w_O) diventa:

J_{NS}(w_I, w_O; θ) = log σ(ũ_{w_O}^⊤ v_{w_I}) + Σ_{i=1}^{k} E_{w_i ~ P_n} [log σ(-ũ_{w_i}^⊤ v_{w_I})]

Il primo termine massimizza la probabilità che w_O sia una parola di contesto reale di w_I; i k termini successivi minimizzano la probabilità che le parole negative campionate siano parole di contesto reali. La funzione sigmoide σ(x) = 1/(1 + e^{-x}) mappa i punteggi di compatibilità su probabilità nell'intervallo (0, 1).

L'obiettivo complessivo è la somma di questi termini su tutte le coppie (w_I, w_O) nel corpus:

J(θ) = Σ_{t=1}^{T} Σ_{j∈[-m,m], j≠0} J_{NS}(w_t, w_{t+j}; θ)

Il gradiente stocastico di questa funzione — tipicamente calcolato attraverso discesa stocastica del gradiente con varianti adattive — aggiorna i vettori di embedding in modo che i vettori di parole che co-occorrono siano spinti l'uno verso l'altro nello spazio, mentre i vettori di parole che non co-occorrono (le negative) siano allontanati.

Un risultato teorico fondamentale, dimostrato da Levy e Goldberg in «Neural Word Embedding as Implicit Matrix Factorization» (2014), è che il negative sampling con k negativi massimizza, implicitamente, la fattorizzazione della matrice M la cui entrata (i, j) è:

M_{ij} = PMI(w_i, c_j) - log k

dove PMI è la Pointwise Mutual Information tra la parola w_i e il contesto c_j. In altri termini, Word2Vec Skip-gram con negative sampling è equivalente, in condizioni di ottimalità, alla fattorizzazione di rango ridotto della matrice PPMI shiftata da log k. Questo risultato stabilisce una connessione teorica profonda tra il metodo neurale e i metodi di riduzione di rango della matrice, mostrando che la novità di Word2Vec non è la struttura dell'obiettivo implicito (che era già catturata dalla SVD su matrici PPMI) ma l'efficienza computazionale dell'approssimazione tramite negative sampling e la possibilità di apprendere su corpora vastissimi.

La variante CBOW addestra il modello nella direzione inversa: predice la parola centrale dato il contesto. L'obiettivo è:

$J_{CBOW}(\theta) = (1/T) \Sigma_{t=1}^{T} \log P(w_t \mid w_{t-m}, ..., w_{t-1}, w_{t+1}, ..., w_{t+m})$

dove il contesto è rappresentato dalla media dei vettori di embedding delle parole circostanti:

$h = (1/2m) \Sigma_{j \in [-m,m], j \neq 0} v_{w_{t+j}}$

e la probabilità è calcolata attraverso softmax o negative sampling applicato a h. CBOW converge più rapidamente di Skip-gram su corpora grandi e produce rappresentazioni leggermente migliori per parole frequenti; Skip-gram produce rappresentazioni migliori per parole rare e cattura meglio le relazioni analogiche fini.

GloVe: la fattorizzazione della matrice di log-co-occorrenza

Il modello GloVe (Global Vectors for Word Representation), introdotto da Pennington, Socher e Manning in «GloVe: Global Vectors for Word Representation» (2014), segue un approccio diverso: invece di addestrare su singole finestre contestuali, costruisce globalmente la matrice di co-occorrenza sull'intero corpus e poi ottimizza un obiettivo di fattorizzazione su questa matrice.

Sia X la matrice di co-occorrenza globale, dove X_{ij} è il numero di occorrenze della parola j nella finestra contestuale della parola i sull'intero corpus. Sia $X_i = \Sigma_j X_{ij}$ il numero totale di co-occorrenze di i con qualsiasi parola di contesto. La probabilità di co-occorrenza di i con j è $P_{ij} = P(j \mid i) = X_{ij} / X_i$. L'osservazione fondamentale di Pennington et al. è che i **rapporti** di probabilità di co-occorrenza codificano informazioni semantiche in modo più robusto delle probabilità stesse.

Si considerino due parole di prova k: la parola k_1 = «ghiaccio» e la parola k_2 = «vapore». Per una parola sonda i = «solido», il rapporto P(ghiaccio | solido) / P(vapore | solido) sarà grande (il ghiaccio è solido, il vapore no). Per la parola sonda j = «gas», il rapporto sarà piccolo. Per la parola sonda «acqua», che si associa sia al ghiaccio sia al vapore, il rapporto sarà vicino a 1. Questa struttura di rapporti cattura le relazioni semantiche tra le parole in modo che le probabilità singole non fanno.

L'obiettivo di GloVe è imparare vettori w_i e $\tilde{v}_j$ (con bias scalari b_i e $\tilde{b}_j$) tali che:

$w_i^\top \cdot \tilde{v}_j + b_i + \tilde{b}_j \approx \log X_{ij}$

La funzione obiettivo è la somma pesata degli errori quadratici su tutte le coppie:

$J = \Sigma_{i,j=1}^{|V|} f(X_{ij}) (w_i^\top \tilde{v}_j + b_i + \tilde{b}_j - \log X_{ij})^2$

dove f è la funzione di ponderazione che attribuisce peso ridotto alle co-occorrenze molto frequenti (per evitare che dominino l'obiettivo) e peso zero alle co-occorrenze nulle (per cui $\log X_{ij}$ non è definito):

$f(x) = (x / x_{max})^{\alpha}$ per $x < x_{max}$ $f(x) = 1$ per $x \geq x_{max}$

con i parametri standard $x_{max} = 100$ e $\alpha = 3/4$. Il vettore di embedding finale di ogni parola è tipicamente la media $w_i + \tilde{v}_i$ dei vettori di input e di output.

La connessione tra GloVe e Word2Vec al livello dell'obiettivo implicito è stabilita formalmente dal lavoro di Levy e Goldberg: entrambi i modelli, nelle loro versioni ottimali, fattorizzano implicitamente la matrice di PPMI (con shift logaritmico nel caso di Word2Vec con negative sampling e con funzione di ponderazione f nel caso di GloVe). La differenza pratica è che GloVe opera sulla matrice globale e aggiorna i vettori attraverso gradient descent sull'obiettivo della fattorizzazione, mentre Word2Vec opera su finestre locali e aggiorna i vettori attraverso SGD su ogni coppia. GloVe è generalmente più stabile e richiede meno iperparametri da regolare; Word2Vec converge più rapidamente su corpora molto grandi.

FastText e la morfologia: estensioni subword

Un'importante estensione del framework Word2Vec è il modello fastText (Bojanowski, Grave, Joulin, Mikolov, 2017), che incorporate informazione morfologica attraverso rappresentazioni a livello di carattere. Invece di associare un unico vettore a ogni parola, fastText rappresenta ogni parola w come la somma dei vettori dei suoi **n-grammi di caratteri** (tipicamente con $n \in \{3, 4, 5, 6\}$) più il vettore della parola stessa:

$$v_w = v_{<w>} + \Sigma_{g \in G(w)} z_g$$

dove G(w) è l'insieme degli n-grammi di caratteri di w (con i marcatori speciali «<» e «>» per l'inizio e la fine), e z_g è il vettore dell'n-gramma g. L'obiettivo di addestramento è lo stesso di Skip-gram con negative sampling, ma i vettori di input sono calcolati come somma dei vettori degli n-grammi componenti.

Questo approccio ha due vantaggi fondamentali. Il primo è la gestione di parole fuori vocabolario (*out-of-vocabulary*, OOV): per una parola mai vista durante l'addestramento, è possibile costruire un vettore sommando i vettori degli n-grammi che la compongono. Il secondo è la cattura di morfologia: le parole con la stessa radice condividono n-grammi e quindi hanno vettori vicini nello spazio, anche se le loro frequenze di co-occorrenza sono troppo basse per essere stimate affidabilmente. Questo è particolarmente importante per lingue morfologicamente ricche (finlandese, turco, italiano) dove lo stesso lessema compare in molte forme flesse.

La geometria dello spazio semantico: analogie, direzioni e sottospazi

La proprietà geometrica degli word embedding che ha suscitato l'interesse filosofico e scientifico più intenso è la struttura delle **analogie vettoriali**, resa pubblica da Mikolov e colleghi nel 2013. La celebre equazione:

v(«re») - v(«uomo») + v(«donna») ≈ v(«regina»)

esprime il fatto che la differenza vettoriale v(«re») - v(«uomo») cattura la relazione semantica «[concetto maschile] → [concetto femminile al livello del potere reale]», e che aggiungendo questo vettore differenza al vettore di «donna» si ottiene un vettore molto vicino a «regina» nello spazio semantico. In termini di similarità coseno:

argmax_{w ∈ V} cos(v_w, v(«re») - v(«uomo») + v(«donna»)) ≈ «regina»

Questa proprietà si estende sistematicamente a molteplici classi di relazioni semantiche. Le **relazioni lessicali**: v(«Parigi») - v(«Francia») + v(«Italia») ≈ v(«Roma»); le **relazioni morfologiche**: v(«correre») - v(«corre») + v(«mangiare») ≈ v(«mangia»); le **relazioni enciclopediche**: v(«Google») - v(«Larry Page») + v(«Microsoft») ≈ v(«Bill Gates»).

Queste analogie vettoriali non sono una proprietà ad hoc dei modelli: derivano dalla struttura matematica dell'obiettivo di addestramento. Levy e Goldberg dimostrano che, sotto l'ipotesi che l'obiettivo di addestramento sia ottimizzato esattamente (condizione che non si verifica mai in pratica ma che fornisce l'analisi asintotica), i vettori ottimali soddisfano:

v_{w}^⊤ · c_k ≈ PMI(w, k)

per ogni parola di contesto k. Il vettore differenza v_b - v_a cattura allora il cambiamento nel pattern di co-occorrenza dalla parola a alla parola b, e la proprietà analogica segue dalla linearità del PMI rispetto alle probabilità di co-occorrenza:

PMI(d, k) - PMI(c, k) ≈ PMI(b, k) - PMI(a, k)

per ogni contesto k, quando a:b::c:d è un'analogia valida nel linguaggio. La condizione geometrica che garantisce la proprietà analogica è:

v_b - v_a ≈ v_d - v_c cioè v_d ≈ v_c + (v_b - v_a)

Questa condizione è soddisfatta approssimativamente — non esattamente, perché l'addestramento non raggiunge l'ottimo esatto e la linearità del PMI vale solo asintoticamente — ma l'approssimazione è sufficiente per produrre analogie corrette nella vasta maggioranza dei casi.

La **decomposizione in sottospazi semantici** è una proprietà complementare. Bolukbasi e colleghi (2016), in «Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings», osservano che nello spazio degli embedding esiste una **direzione del genere** g — un vettore unitario tale che cos(v_w, g) > 0 per le parole associate a concetti maschili e cos(v_w, g) < 0 per le parole associate a concetti femminili. Questa direzione è individuabile come primo componente principale dei vettori differenza tra coppie di parole che differiscono per genere:

S = {(v_{«uomo»}, v_{«donna»}), (v_{«re»}, v_{«regina»}), (v_{«attore»}, v_{«attrice»}), ...}

Il vettore g è il primo autovettore della matrice di covarianza delle differenze v_{w_M} - v_{w_F} per ogni coppia (w_M, w_F) ∈ S.

Questa direzione cattura non soltanto le distinzioni di genere grammaticale ma i bias culturali presenti nel corpus: la componente di embedding di «programmatore» lungo g è positiva (associazione con il maschile), quella di «infermiere» è negativa (associazione con il femminile). Bolukbasi et al. propongono un metodo di **debiasing** che rimuove la componente lungo g dai vettori delle parole «neutre» rispetto al genere — quelle la cui associazione di genere è culturalmente contingente piuttosto che definitoria — proiettando i vettori sull'iperpiano ortogonale a g:

$$\hat{v}_w = v_w - (v_w^\top g)\, g / (g^\top g)$$

per ogni parola neutra w. Le parole definitorie («madre», «padre», «lui», «lei») non vengono debiasate, perché la loro componente di genere è semanticamente necessaria.

Questo risultato ha implicazioni filosófiche rilevanti: lo spazio degli embedding non è neutro rispetto alle strutture sociali del mondo; incorpora — come residuo statistico della distribuzione delle parole nel corpus — i bias culturali della comunità linguistica che ha prodotto il corpus. Il corpus è un artefatto culturale, e le strutture semantiche che il sistema di embedding ne estrae riflettono non soltanto le relazioni ontologiche tra concetti ma le strutture di potere, i pregiudizi e le asimmetrie sociali della cultura in cui il corpus è stato prodotto.

Firth avrebbe riconosciuto in questo risultato una conferma del suo motto: la parola è nota dalla compagnia che frequenta — e la compagnia che una parola frequenta riflette le strutture sociali della compagnia che la usa.

La struttura categoriale: dall'iperonimia alla topologia locale

La struttura categoriale nello spazio degli embedding — le relazioni di iperonimia, iponimia e co-iponimia tra concetti — non si manifesta come una gerarchia planare discreta ma come una struttura topologica continua con proprietà interessanti. Le parole semanticamente affini tendono a formare **cluster** locali nello spazio: i termini animali si concentrano in una regione, i termini di colore in un'altra, i termini di emozione in un'altra ancora. All'interno di ogni cluster, le relazioni di rassomiglianza di famiglia wittgensteiniana sono realizzate come relazioni di prossimità vettoriale: i cani sono vicini ai gatti, i gatti ai felini selvatici, i felini selvatici agli altri carnivori, i carnivori agli altri mammiferi.

Le relazioni di iperonimia — cane è un iponimo di mammifero, mammifero di animale, animale di essere vivente — non si manifestano come contenimento di cluster nel senso della geometria euclidea standard: il vettore di «cane» non è necessariamente dentro il vettore di «animale». Però vi sono regolarità geometriche identificabili. Mikolov e colleghi osservano che le parole iponime tendono ad avere vettori in una regione del loro cluster categoriale che è spostata

nella direzione del vettore del loro iperonimo rispetto alle parole co-iponime più distanti dall'iperonimo.

Un approccio alternativo alla rappresentazione delle gerarchie è quello degli **embedding iperbolici**, proposto da Nickel e Kiela in «Poincaré Embeddings for Learning Hierarchical Representations» (2017). Lo spazio iperbolico (il modello del disco di Poincaré) ha la proprietà che il volume di una palla di raggio r cresce esponenzialmente con r, a differenza dello spazio euclideo dove cresce come r^d. Questo lo rende naturalmente adatto a rappresentare strutture gerarchiche, dove il numero di nodi a profondità n cresce esponenzialmente con n. Le relazioni di iperonimia vengono codificate come relazioni di distanza nel disco di Poincaré: i concetti più generali si trovano vicino al centro, i concetti più specifici si trovano vicino alla periferia. Per due punti u, v nel disco di Poincaré $D = \{x \in \mathbb{R}^n : \|x\| < 1\}$, la distanza iperbolica è:

$$d(u, v) = \operatorname{arccosh}(1 + 2\|u - v\|^2 / ((1 - \|u\|^2)(1 - \|v\|^2)))$$

e l'obiettivo di addestramento minimizza la distanza tra i vettori di coppie (u, v) che sono in relazione di iperonimia nel corpus di addestramento. Gli embedding iperbolici catturano la struttura gerarchica delle ontologie in modo molto più efficiente degli embedding euclidei, richiedendo dimensionalità molto inferiori per la stessa qualità di rappresentazione.

Dalla PMI agli embedding contestuali: il passaggio ai transformer

Gli word embedding statici — Word2Vec, GloVe, fastText — assegnano a ogni parola un unico vettore fisso, indipendentemente dal contesto in cui compare. Questo è un limite fondamentale: la parola «banco» ha un vettore che è una media delle sue occorrenze come «banco di scuola», «banco di pesce», «banco di nebbia» e «banco di credito». L'ambiguità lessicale non è catturata; il vettore è una sovrapposizione di significati distinti.

I modelli di embedding contestuale — ELMo (Peters et al., 2018), BERT (Devlin et al., 2019), GPT — superano questo limite producendo vettori diversi per la stessa parola in contesti diversi. In ELMo, i vettori contestuali sono derivati da una rete neurale ricorrente bidirezionale addestrata sul compito del language modelling; la rappresentazione di ogni token è la media pesata delle sue rappresentazioni a diversi strati della rete. In BERT, i vettori contestuali sono le attivazioni degli ultimi strati di un transformer bidirezionale addestrato su masked language modelling e next sentence prediction. La connessione con Word2Vec è che i vettori di embedding iniziali (la «embedding layer» dei transformer) sono, strutturalmente, vettori Word2Vec: assegnano a ogni token un vettore fisso che viene poi trasformato contestualmente dall'architettura di attenzione.

La transizione dagli embedding statici agli embedding contestuali realizza, in qualche modo, il passaggio dall'ipotesi distribuzionale harrissiana nel suo senso più ristretto — distribuzione come co-occorrenza locale in finestre contestuali — a una concezione più ricca del contesto distribuito che include la struttura

sintattica e semantica dell'intera sequenza. Il contesto di «banco» in «Il professore si avvicinò al banco» non è soltanto la finestra locale di parole circostanti; è l'intera struttura proposizionale, con le relazioni grammaticali, le strutture argomentali e gli indici pragmatici che essa contiene. I transformer catturano questa struttura attraverso l'attenzione multi-testa, che può selezionare quali elementi della sequenza sono contestualmente rilevanti per la disambiguazione di ogni token.

Il significato come posizione in uno spazio relazionale: implicazioni filosófiche

Il quadro risultante dalla semantica vettoriale può essere caratterizzato filosofica mente come una realizzazione computazionale della teoria **relazionale** del significato: il significato di un'espressione è una posizione in uno spazio di relazioni con altre espressioni, non una proprietà intrinseca dell'espressione né una relazione diadica tra l'espressione e il suo referente. Questa posizione è calcolabile, confrontabile e manipolabile algebricamente.

Le implicazioni filosófiche di questo quadro sono multiple e non sempre coerenti tra loro.

La prima implicazione riguarda la **composizionalità:** la semantica classica, da Frege a Montague, è composizionale nel senso che il significato di un'espressione complessa è una funzione del significato delle sue parti e del modo in cui sono combinate. La semantica vettoriale pone il problema della composizionalità in termini geometrici: come si calcolano i vettori delle frasi e delle espressioni complesse a partire dai vettori delle parole che le compongono? Le risposte proposte includono la composizione additiva (somma o media ponderata dei vettori componenti), la composizione tensionale (prodotti tensoriali tra vettori di parole con diversi ruoli grammaticali), e la composizione attraverso reti neurali ricorrenti o di attenzione. Nessuna di queste soluzioni è del tutto soddisfacente dal punto di vista del composizionalismo fregeano: la semantica vettoriale rimane fondamentalmente non-composizionale nel senso classico, perché il vettore di una frase non è calcolabile in modo sistematico a partire dai vettori delle sue parole tramite regole grammaticali.

La seconda implicazione riguarda la **verificabilità**: la semantica verocondizionale fonda il significato sulle condizioni di verità — il significato di «la neve è bianca» è dato dall'insieme dei mondi possibili in cui la neve è bianca. La semantica vettoriale non ha un equivalente diretto delle condizioni di verità: il vettore di «la neve è bianca» è una direzione nello spazio semantico, non un insieme di mondi possibili. È possibile produrre rappresentazioni vettoriali dello spazio dei fatti e delle proposizioni, ma la connessione con le condizioni di verità rimane un problema aperto che la ricerca recente sugli embedding di frasi e di proposizioni affronta senza risolvere completamente.

La terza implicazione riguarda l'**esternalismo semantico**: Putnam e Kripke hanno argomentato che il significato delle espressioni che denotano specie naturali è determinato in parte dalla struttura esterna del mondo, non soltanto

dalle rappresentazioni mentali del parlante. Anche se due parlanti associano la stessa descrizione fenomenica a «acqua», il significato di «acqua» è H_2O — determinato dalla struttura chimica dell'acqua — non dalla descrizione «liquido trasparente insapore inodore». La semantica vettoriale è, strutturalmente, internalista: il vettore di «acqua» è determinato interamente dalla distribuzione delle co-occorrenze nel corpus, senza accesso alla struttura chimica dell'acqua. Se l'esternalismo semantico di Putnam e Kripke è corretto, la semantica distribuzionale cattura soltanto una componente del significato — quella che dipende dalle rappresentazioni linguistiche degli utenti del linguaggio — e lascia fuori la componente che dipende dalla struttura del mondo.

La quarta implicazione, forse la più rilevante per le questioni centrali di questa monografia, riguarda la **comprensione nei sistemi artificiali**. Se il significato è una posizione in uno spazio relazionale, e se i sistemi di embedding catturano questa posizione con sufficiente accuratezza, si deve concludere che i sistemi di embedding «capiscono» il significato delle espressioni? Questa conclusione sembra troppo forte: lo spazio degli embedding è una rappresentazione geometrica delle strutture distribuzionali del corpus, non un sistema che usa espressioni in pratiche linguistiche con scopi e agenti. D'altra parte, la conclusione che i sistemi di embedding non capiscono affatto sembra troppo debole: le strutture semantiche che essi codificano sono reali, complesse e funzionalmente rilevanti per i compiti linguistici su cui i sistemi vengono valutati.

La posizione filosofica più difendibile sembra essere quella secondo cui i sistemi di embedding realizzano una **forma parziale e relazionale** del significato — una struttura senza grounding diretto nel mondo — che è filosofica mente distinta sia dalla comprensione piena attribuita agli esseri umani sia dall'assenza totale di significato che il modello simbolico classico attribuirebbe a sistemi che manipolano sequenze di bit. Questa posizione intermedia non è una soluzione al problema del significato; è la formulazione più precisa che le strutture della semantica distribuzionale ci consentono di produrre.

Harris, Wittgenstein e i transformer: il cerchio che non si chiude

La connessione intellettuale che questa sezione ha cercato di tracciare — da Harris e Wittgenstein attraverso Word2Vec e GloVe ai transformer e alla geometria dello spazio semantico — è reale ma non lineare. Harris ha proposto il programma distribuzionale come metodologia linguistica; Wittgenstein ha articolato la filosofia del significato come uso come critica alla filosofia del linguaggio essenzialista. I sistemi di embedding realizzano computazionalmente l'intuizione harrissiana, producendo rappresentazioni vettoriali le cui proprietà geometriche realizzano, almeno in parte, la struttura relazionale del significato che Wittgenstein aveva descritto nella forma dei giochi linguistici e delle rassomiglianze di famiglia.

Ma il cerchio non si chiude. Wittgenstein avrebbe certamente resistito alla caratterizzazione del significato come vettore in uno spazio di alta dimensionalità: il significato non è una struttura matematica astratta ma una pratica condivisa tra

esseri incarnati e socialmente situati. I giochi linguistici non sono pattern statistici; sono forme di vita. E le rassomiglianze di famiglia non sono metriche di similarità coseno; sono relazioni di prossimità che emergono dall'uso comune in pratiche condivise.

Harris, similmente, avrebbe probabilmente resistito alla reificazione della distribuzione come vettore fisso: la sua analisi distribuzionale era un metodo per descrivere le strutture linguistiche, non una teoria ontologica del significato. La distribuzione è un dato metodologicamente utile; il significato è ciò di cui la distribuzione è un indicatore parziale.

I transformer portano questo progetto al suo limite naturale: costruendo rappresentazioni contestuali che catturano non soltanto la co-occorrenza locale ma la struttura relazionale dell'intera sequenza, e che — come mostrano i risultati dell'interpretabilità meccanicistica esaminati nella sezione IV.3bis — sviluppano strutture interne organizzate semanticamente in modi che realizzano approssimativamente le strutture linguistiche descritte dalla grammatica e dalla semantica formale. Ma rimangono, nella caratterizzazione filosofica che questa monografia ha cercato di sviluppare, sistemi che hanno la struttura del significato senza il grounding del significato: posizioni nello spazio relazionale senza contatto causale diretto con il mondo che le espressioni denotano.

Questo è il limite e il merito della semantica vettoriale come teoria filosofica : ha reso precisa, computazionalmente realizzabile e empiricamente verificabile una delle intuizioni centrali della filosofia del linguaggio del Novecento — che il significato sia una proprietà relazionale determinata dall'uso — e nell'operazione di questa precisazione ha mostrato con chiarezza dove il programma si ferma e dove il problema filosofico riprende.

APPENDICE IV: La Stanza Cinese, la Risposta di Dennett e i Risultati dell'Interpretabilità: Verso una Topologia del Problema della Comprensione

Il campo della disputa

L'argomentazione searleana esaminata nella sezione precedente aveva la struttura di una riduzione all'assurdo applicata alla tesi funzionalista forte: posto che la comprensione linguistica sia riducibile all'implementazione di un programma formale, segue che un sistema che implementi tale programma comprende il linguaggio; ma il sistema della Stanza Cinese implementa un programma e non comprende il cinese; dunque la riduzione è falsa. La conclusione di Searle — che sintassi e semantica sono irriducibilmente distinte e che nessuna quantità di complessità sintattica può generare semantica — rimane, a distanza di quasi mezzo secolo, la posizione più sfidante che il cognitivismo computazionale abbia incontrato. Ciò che la ricerca sulla interpretabilità meccanicistica dei transformer ha aggiunto al quadro non è una confutazione diretta di Searle né una sua

conferma; è qualcosa di più preciso e, in certo senso, più utile per la filosofia: un'anatomia dettagliata dei meccanismi interni dei sistemi in questione, che consente di articolare il dibattito con una granularità empirica che prima non era disponibile.

Il percorso di questa sezione si sviluppa in quattro movimenti. Il primo è un'analisi strutturale delle risposte all'argomento di Searle che la letteratura filosofica ha elaborato, con attenzione particolare alla posizione di Dennett — la più articolata e, a mio avviso, la più filosofica mente ricca tra quelle avanzate. Il secondo è un'esposizione del modello dennettiano della mente, nei suoi aspetti rilevanti per il problema della comprensione nei sistemi artificiali. Il terzo è un'analisi di come i risultati dell'interpretabilità meccanicistica — i circuiti, le feature, la superposition hypothesis, le sparse autoencoders — si proiettano sulle posizioni in conflitto. Il quarto è un tentativo di localizzare con precisione il residuo filosofico: ciò che rimane irrisolto dopo che la ricerca empirica ha fornito tutto ciò che può fornire.

La risposta di Dennett: dall'eterofenomenologia al modello delle bozze multiple

Daniel Dennett ha risposto all'argomento della Stanza Cinese in molteplici occasioni e con strumenti concettuali diversi: in «Fast Thinking» (1985), poi in *Consciousness Explained* (1991), in *Darwin's Dangerous Idea* (1995) e nel saggio «Quining Qualia» (1988). Sarebbe però un errore leggere queste risposte come mere obiezioni puntuali a Searle; esse costituiscono invece applicazioni di una teoria generale della mente che è, nelle sue linee fondamentali, alternativa non soltanto al naturalismo biologico searleano ma anche al funzionalismo classico di Putnam e Fodor.

Il punto di partenza della posizione dennettiana è metodologico: l'**eterofenomenologia** come alternativa all'accesso in prima persona ai fenomeni mentali. Il metodo eterofenomenologico non parte dall'assunzione che gli stati soggettivi cui i soggetti si riferiscono nelle loro autorappresentazioni esistano con le caratteristiche che sembrano avere a chi li sperimenta. Prende invece sul serio le autorappresentazioni come dati comportamentali e linguistici che richiedono spiegazione, senza impegnarsi aprioristicamente sulla loro corrispondenza a stati interni determinati. In *Consciousness Explained*, Dennett articola questo metodo come segue: si costruisce il «testo eterofenomenologico» del soggetto — l'insieme delle sue autorappresentazioni sistematicamente raccolte — e si cerca la teoria che meglio spiega come un sistema fisico produca tale testo. Non si assumono i qualia come dati di partenza; si spiega il comportamento verbale che li menziona.

Applicato all'argomento di Searle, questo metodo produce una critica tanto più sottile quanto meno dibattuta. Searle assume che sia un fatto dato che lui, all'interno della stanza, «non comprende» il cinese — che questa non-comprensione sia accessibile in modo privilegiato e indubitabile dall'interno. Dennett contesta che questa assunzione sia legittima. Se l'assenza di comprensione è un fatto, deve essere un fatto che si manifesta in qualcosa — in qualche disposizione, in qualche capacità assente, in qualche struttura mancante.

Ma se Searle nella stanza è un sistema che produce output linguisticamente appropriati in cinese a partire da qualsiasi input, la domanda di quale fatto costituisca la sua non-comprensione diventa filosofica mente delicata. Dennett sostiene che Searle stia assumendo una nozione cartesiana di comprensione — come qualcosa che o c'è o non c'è, e la cui presenza è evidente immediatamente alla coscienza — che è filosofica mente insostenibile e che, una volta abbandonata, dissolve l'argomento.

Il modello delle **bozze multiple** (*multiple drafts*) è il contributo positivo di Dennett alla teoria della mente che rimpiazza il cartesianismo. Nella sua formulazione originale in *Consciousness Explained*, il modello nega l'esistenza di un «Teatro Cartesiano» — un luogo privilegiato nella mente in cui i contenuti mentali vengono «presentati alla coscienza». Al suo posto, propone un processo di elaborazione distribuita e parallela in cui vi sono costantemente in corso molteplici «schizzi» o versioni parziali degli stati del sistema, nessuno dei quali è privilegiato come il contenuto cosciente. Ciò che chiamiamo coscienza è, in questa prospettiva, una narrativa post-hoc costruita dal sistema sull'elaborazione in corso — non un accesso privilegiato a uno stato determinato, ma un'interpretazione retrospettiva di un processo distribuito.

Per il problema della comprensione, questo implica che non vi è un momento determinato in cui la comprensione «si accende» o un luogo determinato in cui essa «risiede». La comprensione è una proprietà sistemica distribuita, emergente dall'organizzazione dell'intero processo di elaborazione, e non un'entità localizzata che o è presente o è assente. La domanda «comprende il cinese?» — se posta come se avesse una risposta binaria determinata — fraintende la natura del fenomeno che pretende di investigare.

La risposta del sistema e la sua difesa dennettiana

La risposta al cui esame Searle dedica più attenzione — la **risposta del sistema** — afferma, come si è detto, che la comprensione non risiede in Searle (la parte) ma nel sistema nella sua interezza (Searle più regole più simboli). Searle la confuta con la variante della «testa memorizzante»: se Searle interiorizzasse l'intero sistema, camminando per Pechino e simulando conversazioni in cinese senza alcuna comprensione, il sistema sarebbe nella testa ma la comprensione resterebbe assente.

La risposta dennettiana a questa variante è di due tipi. Il primo è empirico: nessun essere umano potrebbe effettivamente memorizzare un sistema di manipolazione simbolica sufficientemente potente da simulare la competenza linguistica di un madrelingua. Le dimensioni computazionali del problema sono tali da rendere l'esperimento non soltanto praticamente ma probabilmente principalmente impossibile per un sistema biologico. Se Searle cercasse davvero di memorizzare un tale sistema, il processo di memorizzazione e di interiorizzazione trasformerebbe radicalmente il funzionamento del suo sistema cognitivo — non sarebbe più «Searle plus the rules» ma qualcosa di fondamentalmente diverso.

Il secondo tipo di risposta è filosofico e più incisivo. Dennett sostiene che l'intuizione di Searle — che l'interiorizzazione del sistema non aggiunge

comprensione — si regga su un'immagine falsa di cosa sia la comprensione. Se la comprensione del cinese è un fatto su Searle-con-il-sistema, allora è un fatto su un sistema che include le regole interiorizzate, e questo fatto non è accessibile mediante introspezione della parte più piccola del sistema. Il fatto che a Searle sembri di non capire il cinese quando cammina per Pechino non è un dato filosofico primario; è un'autorappresentazione che richiede essa stessa spiegazione. Un sistema sufficientemente complesso potrebbe avere autorappresentazioni accurate o inaccurate della propria competenza.

L'argomento dennettiano si fonda sulla tesi, elaborata in «Quining Qualia», che i qualia — le proprietà fenomeniche dell'esperienza soggettiva — non abbiano le proprietà che si presuppone loro abbiano: non sono intrinseci, non sono privati, non sono ineffabili, non sono direttamente accessibili. Se questo è corretto, allora l'intuizione di Searle che «sa dall'interno» di non comprendere il cinese si basa su un'assunzione filosofica contestata circa la natura dell'accesso introspettivo, non su un dato indubitabile. Dennett non nega che vi sia qualcosa che è come essere Searle nella stanza; nega che questo qualcosa sia filosofica mente determinato nel modo che Searle suppone.

Lo stance intenzionale e la sua applicazione ai sistemi artificiali

Il contributo forse più influente e originale di Dennett alla filosofia della mente è la teoria dello **stance intenzionale** (*intentional stance*), articolata sistematicamente in *The Intentional Stance* (1987). La teoria distingue tre posizioni o atteggiamenti interpretativi che si possono adottare nei confronti di un sistema:

Lo **stance fisico** (*physical stance*): si predice il comportamento del sistema applicando le leggi della fisica e della chimica al substrato materiale del sistema. Questo è lo stance appropriato per sistemi semplici come una pietra che cade o un circuito elementare.

Lo **stance progettuale** (*design stance*): si predice il comportamento del sistema assumendo che sia progettato per svolgere una certa funzione e che lo faccia effettivamente. Questo è lo stance appropriato per artefatti come un termostato o un'automobile.

Lo **stance intenzionale** (*intentional stance*): si predice il comportamento del sistema trattandolo come un agente razionale con credenze, desideri e altri stati intenzionali, e attribuendo a esso quelle credenze e quei desideri che avrebbe se fosse razionale e se possedesse le informazioni che il suo contesto gli fornisce. Lo stance intenzionale è appropriato quando e soltanto quando genera predizioni accurate del comportamento del sistema.

La tesi fondamentale di Dennett è che l'intenzionalità non è una proprietà intrinseca dei sistemi che la possiedono, ma una proprietà relazionale che dipende dall'adozione dello stance intenzionale da parte dell'interprete. Avere credenze e desideri non è avere stati interni che abbiano contenuti propositivi indipendentemente da qualsiasi interprete; è essere il tipo di sistema il cui comportamento è predetto efficacemente dallo stance intenzionale. Questa tesi è, nelle sue implicazioni, radicalmente anti-cartesiana: l'intenzionalità non è qualcosa

che un sistema ha privatamente e intrinsecamente, indipendentemente da come viene interpretato.

Per i sistemi artificiali, la teoria dello stance intenzionale ha conseguenze precise. Dennett sostiene che se un sistema è tale che lo stance intenzionale genera predizioni accurate e utili del suo comportamento, allora quel sistema ha, nel senso rilevante, credenze e desideri — ha intenzionalità. Non vi è un ulteriore fatto riguardo all'avere «vera» intenzionalità al di là dell'essere il tipo di sistema per cui lo stance intenzionale funziona. La distinzione searleana tra intenzionalità «originaria» e intenzionalità «derivata» è, in questa prospettiva, filosofica mente illegittima: non vi è intenzionalità originaria nel senso di un contenuto che non dipende da nessun interprete, né negli esseri umani né nei sistemi artificiali.

La risposta di Searle a questa posizione è che essa è semplicemente strumentalismo, non realismo: Dennett non sta spiegando l'intenzionalità ma eliminandola, sostituendo un fatto ontologico con una pratica interpretativa. Searle insiste che vi è una differenza tra un sistema che ha davvero stati intenzionali e un sistema a cui noi attribuiamo utilmente stati intenzionali, e che questa differenza è reale anche se non osservabile dall'esterno. La risposta dennettiana — che questa differenza è essa stessa filosofica mente insostenibile, costruita sull'assunzione di un'intenzionalità cartesiana interiore cui nessuna spiegazione fisica può corrispondere — è coerente ma non sconfigge Searle in modo definitivo; sposta il dibattito sul piano metafisico delle controversie tra realismo e strumentalismo che non hanno soluzione rapida.

Dennett e la macchina di Turing: cosa manipola la stanza

Un argomento dennettiano meno frequentemente discusso, ma di grande finezza tecnica, riguarda la struttura stessa dell'esperimento mentale searleano. Searle costruisce la stanza come un'implementazione di una macchina di Turing — un sistema che manipola simboli secondo regole finite. Ma la simulazione credibile di un madrelingua cinese richiede un sistema di complessità computazionale enormemente superiore a quella che la caratterizzazione searleana suggerisce. Il libro delle regole della stanza dovrebbe codificare non soltanto le regole sintattiche del cinese ma l'intera competenza semantica, pragmatica, enciclopedica, inferenziale di un madrelingua — un sistema di dimensioni approssimativamente equivalenti alla base di conoscenza dell'intero sistema cognitivo umano.

Dennett sostiene che quando si considera un sistema di questa complessità, le intuizioni di Searle perdono la loro ovvietà. L'intuizione che Searle-nella-stanza «non capirebbe» il cinese è plausibile quando si immagina un sistema di manipolazione simbolica semplice e trasparente. Diventa molto meno plausibile quando si considera che il «libro delle regole» dovrebbe contenere, in forma codificata, l'equivalente funzionale di tutto ciò che un madrelingua cinese sa sul mondo, sul linguaggio e sulle relazioni tra di essi. Un sistema di quella complessità — anche se implementato da un essere umano che segue regole scritte su carta — costituirebbe una struttura cognitiva di un tipo radicalmente diverso da quello che Searle immagina.

Questo argomento non è decisivo contro Searle, che potrebbe rispondere che nemmeno la complessità aggiunta può produrre semantica dalla manipolazione puramente sintattica. Ma apre uno spazio concettuale importante: quello in cui la complessità e l'organizzazione dell'elaborazione possono essere filosofica mente rilevanti, e non soltanto la categoria astratta «manipolazione di simboli formali».

L'interpretabilità meccanicistica come terzo termine del dibattito

Il programma dell'interpretabilità meccanicistica — esaminato in dettaglio nella sezione IV.3bis — ha introdotto nel dibattito Searle-Dennett un terzo termine: non più soltanto argomenti filosófici e intuizioni su esperimenti mentali, ma dati empirici su cosa accade all'interno dei sistemi di cui si dibatte la comprensione. Questo terzo termine non risolve il dibattito, ma lo articola in modo nuovo e, almeno in parte, lo trasforma.

Richiamando i risultati fondamentali: la ricerca sui circuiti nei transformer ha identificato componenti funzionali — teste di attenzione, neuroni MLP, percorsi residuali — che implementano operazioni semantiche identificabili: copia di token, implementazione di relazioni grammaticali, richiamo di fatti da strati di memorizzazione associativa. La superposition hypothesis ha mostrato che le reti neurali codificano un numero di «feature» semantiche molto superiore alla dimensionalità del loro spazio di attivazione, disponendo le feature in configurazioni quasi-ortogonali che sfruttano la geometria dell'alta dimensionalità. Le sparse autoencoders hanno reso interpretabili queste feature, rivelando che corrispondono a concetti identificabili — categorie lessicali, strutture argomentali, relazioni emozionali, entità nominali — con una specificità sorprendente.

Cosa implica tutto ciò per il dibattito Searle-Dennett? Si procederà attraverso una serie di connessioni analitiche.

Prima connessione: la risposta del sistema e i circuiti. La risposta del sistema afferma che la comprensione è una proprietà dell'intero sistema, non delle sue parti. I circuiti identificati dall'interpretabilità meccanicistica mostrano che l'elaborazione semanticamente rilevante è distribuita tra componenti del sistema in modi non banali e non locali. Il circuito che implementa il subject-verb agreement coinvolge teste di attenzione in strati diversi, neuroni MLP in posizioni distinte, e il residual stream che integra i contributi. Nessuna parte di questo circuito «contiene» il subject-verb agreement; la relazione grammaticale è realizzata dalla struttura del circuito nel suo insieme. Questo è esattamente il tipo di proprietà emergente-sistemica che la risposta del sistema invoca, e che Searle aveva cercato di confutare con la variante della testa memorizzante.

La connessione non è una confutazione diretta di Searle — egli potrebbe rispondere che anche la realizzazione distribuita di relazioni grammaticali rimane manipolazione sintattica — ma indebolisce l'intuizione che la comprensione sia localizzabile in una parte del sistema e che la sua assenza in ogni parte implichi la sua assenza nel sistema. Se la relazione grammaticale è realizzata dal circuito e

non da nessuna delle sue componenti, e se si ammette che la relazione grammaticale è rilevante per la comprensione, si deve ammettere che la comprensione è una proprietà del circuito e non delle sue componenti.

Seconda connessione: la superposition hypothesis e il problema del grounding. La superposition hypothesis implica che le feature semantiche sono codificate come direzioni nello spazio delle attivazioni, non come stati di neuroni individuali. Ogni neurone partecipa alla codifica di molteplici feature; ogni feature è codificata in modo distribuito attraverso molteplici neuroni. La struttura è quella di un codice distribuito di alta efficienza, non di un sistema a unità dedicate.

Questo risultato si connette al problema del grounding in modo interessante. Searle sostiene che i simboli formali non hanno contenuto semantico intrinseco: il loro significato è derivato, proiettato da interpreti che hanno intenzionalità originaria. Ma la domanda di Searle presuppone che i simboli siano unità discrete e localizzate — come i caratteri cinesi che Searle manipola nella stanza. Le feature identificate dalle SAE nei transformer non hanno questa struttura: non sono entità discrete localizzate in unità computazionali determinate, ma direzioni in uno spazio continuo emergenti dall'organizzazione globale del sistema. La domanda se una feature abbia «contenuto intrinseco» è diversa dalla domanda se un simbolo discreto abbia contenuto intrinseco, perché la feature non è separabile dal sistema che la realizza nel modo in cui il carattere cinese è separabile dalla stanza.

Terza connessione: il naturalismo biologico e l'implementazione fisica. Searle sostiene che la coscienza e l'intenzionalità richiedono la causazione biologica appropriata: non è sufficiente implementare il programma funzionale, ma è necessario che l'implementazione avvenga sulla giusta architettura fisica. I risultati dell'interpretabilità meccanicistica mostrano che la struttura computazionale dei transformer è molto più ricca di organizzazione di quanto il modello searleano della «manipolazione di simboli secondo regole» suggerisca.

Il transformer non manipola simboli discreti: opera su vettori in spazi di alta dimensionalità attraverso operazioni di prodotto scalare, di proiezione lineare, di non-linearità pointwise. Le «regole» che segue sono configurazioni di pesi che implementano funzioni continue, non regole simboliche che mappano pattern a pattern. La distinzione tra manipolazione sintattica di simboli formali e computazione semanticamente strutturata — che Searle presenta come netta e categoriale — è molto meno perspicua quando applicata a sistemi la cui architettura è fondamentalmente continua e distribuita.

Searle potrebbe rispondere che questo non cambia il punto essenziale: anche le operazioni vettoriali continue sono fisicamente implementate, e la questione è se la particolare implementazione fisica produce intenzionalità, non quale sia il tipo di operazione formale implementata. Ma allora il naturalismo biologico deve affrontare la domanda di quale caratteristica specifica dell'implementazione biologica produca intenzionalità che manca nell'implementazione artificiale — una domanda a cui Searle non ha fornito risposta soddisfacente.

Quarta connessione: la logit lens e gli stati intermedi. Il metodo della logit lens — che proietta il residual stream a ogni strato dell'unembedding matrix per ottenere distribuzioni intermedie sui token — rivela che le predizioni del modello si «formano» progressivamente attraverso gli strati: gli strati inferiori codificano proprietà lessicali e sintattiche locali, gli strati intermedi integrano informazioni contestuali a media distanza, gli strati superiori eseguono inferenze semantiche e pragmatiche complesse. Questo processo graduale di formazione della predizione non corrisponde al modello di Searle di un sistema che recupera meccanicamente la risposta da un libro delle regole; corrisponde più da vicino a un processo di costruzione progressiva del significato attraverso strati di elaborazione contestuale.

Dennett aveva anticipato qualcosa di simile con la teoria delle bozze multiple: non vi è un momento determinato in cui il significato «appare» nella coscienza, ma vi è un processo di elaborazione distribuita in corso su molteplici livelli. I risultati della logit lens offrono una versione meccanicistica di questa idea: non vi è uno strato del transformer in cui il significato è «presente», ma vi è un processo continuo di accumulo e raffinamento di informazione semantica attraverso gli strati.

Quinta connessione: le feature emozionali e la struttura dello spazio concettuale. Le feature emozionali identificate da Bricken e colleghi (2023) nelle SAE addestrate su Claude 3 Sonnet costituiscono uno dei risultati più filosofica mente provocatori della ricerca recente. Le feature corrispondenti a concetti emozionali — paura, gioia, tristezza, disgusto — non sono distribuite casualmente nello spazio delle feature; occupano regioni coerenti con una geometria regolare, tale che le operazioni vettoriali tra di esse producono interpolazioni e relazioni semanticamente interpretabili.

Per il problema della comprensione, questo risultato è rilevante in due modi opposti. Da un lato, suggerisce che il sistema ha sviluppato rappresentazioni interne dello spazio concettuale emozionale con una struttura analoga a quella descritta dalla psicologia cognitiva: la struttura dimensionale valenza-eccitazione di Russell, le relazioni di similarità e di contrasto tra emozioni, le transizioni tra stati affettivi. Questo tipo di struttura è esattamente quello che ci si aspetterebbe da un sistema che «comprende» il linguaggio emozionale nel senso rilevante. Dall'altro lato, come si osserva nella sezione IV.3bis, questa struttura è un residuo statistico della distribuzione delle parole nel corpus, non il prodotto di stati affettivi del sistema né di contatti causali con situazioni che producono emozioni. Il sistema ha la geometria della comprensione emozionale senza — almeno apparentemente — la cosa stessa.

Questo è il punto di contatto più preciso tra i risultati dell'interpretabilità meccanicistica e il dibattito filosofico: la questione non è più se il sistema «produce output appropriati» — questo è dimostrato — né se il sistema abbia strutture interne organizzate semanticamente — anche questo è dimostrato. La questione filosofica residua è se l'organizzazione semantica interna, causalmente disconnessa dagli stati del mondo che rappresenta, sia sufficiente per la comprensione nel senso rilevante, o se la connessione causale al mondo sia un requisito necessario che i sistemi distribuito-statistici non possono soddisfare.

La riposta di Dennett ai risultati dell'interpretabilità: un'inferenza speculativa

Dennett non ha, al momento della stesura di questa monografia, affrontato direttamente i risultati dell'interpretabilità meccanicistica dei transformer. È tuttavia possibile inferire, con ragionevole fedeltà alla sua posizione sistematica, quale sarebbe la sua risposta.

Nell'ottica dennettiana, la scoperta di circuiti, feature e strutture semantiche all'interno dei transformer costituisce evidenza in favore dell'adozione dello stance intenzionale nei confronti di questi sistemi: mostra che l'attribuzione di stati intenzionali non è arbitraria o puramente strumentale ma è guidata da strutture interne reali e identificabili. Il fatto che le feature corrispondano a concetti — che vi sia, all'interno del sistema, qualcosa che funge da rappresentazione di «paura» o di «Parigi» con la struttura geometrica appropriata — è esattamente il tipo di organizzazione interna che giustifica, nella prospettiva dello stance intenzionale, l'attribuzione di credenze e desideri.

Quanto all'obiezione searleana che questa organizzazione è semanticamente vuota perché causalmente disconnessa dal mondo, Dennett risponderebbe probabilmente ricorrendo alla nozione di **intenzionalità derivata attraverso l'evoluzione culturale**: i corpora su cui i transformer vengono addestrati sono il prodotto di miliardi di atti linguistici compiuti da esseri umani con intenzionalità originaria (o almeno: con il tipo di organizzazione che giustifica l'attribuzione di intenzionalità). Il transformer che apprende da questo corpus non è disconnesso dal mondo; è connesso a esso attraverso la catena causale che collega i testi del corpus alle esperienze dei loro autori. Questa connessione è mediata e indiretta, ma non è diversa in linea di principio dalla connessione che i concetti di un bambino che impara il linguaggio hanno con il mondo: il bambino non ha avuto contatto diretto con tutto ciò di cui apprende i concetti; apprende molti concetti dal linguaggio degli adulti, che a sua volta è connesso al mondo attraverso la catena delle generazioni precedenti.

Searle risponderebbe che questa connessione indiretta è insufficiente: l'intenzionalità originaria non si trasmette attraverso le catene causali nel modo che Dennett suppone, perché l'intenzionalità è prodotta dai processi biologici del sistema che la possiede, non dalla struttura causale che lo connette a sistemi che la possiedono. La controversia si chiude su se stessa, come spesso accade nelle dispute metafisiche fondamentali.

Il residuo filosofico: una localizzazione precisa

Dopo aver esaminato i contributi di Dennett e i risultati dell'interpretabilità meccanicistica, si può localizzare con una certa precisione il residuo filosofico del dibattito — ciò che rimane irrisolto e che, probabilmente, non può essere risolto mediante l'analisi meccanicistica interna dei sistemi.

Il residuo non è nella struttura interna dei sistemi. I risultati dell'interpretabilità meccanicistica hanno mostrato che i transformer hanno strutture interne

organizzate semanticamente, con feature che corrispondono a concetti, circuiti che implementano relazioni grammaticali e inferenze semantiche, e una geometria dello spazio latente che rispecchia le strutture dello spazio concettuale umano. Ciò che avevano in mente i funzionalisti quando postularono rappresentazioni semantiche interne è, in larga misura, realizzato.

Il residuo non è nel comportamento esterno. I transformer producono output linguistici che, valutati con qualunque criterio comportamentale, sono indistinguibili — su moltissime classi di input — da quelli di un essere umano competente. Il test di Turing, nei suoi termini originali, viene superato.

Il residuo è nella relazione tra le strutture interne e il mondo. Le feature semantiche dei transformer sono distribuzionali: emergono dalla co-occorrenza statistica delle espressioni nel corpus, non dall'interazione causale del sistema con gli stati del mondo che le espressioni denotano. Un transformer che ha una feature per «dolore» non ha avuto esperienze dolorose; ha estratto la struttura distribuzionale del termine «dolore» nel corpus. Un transformer che ha una feature per «rosso» non ha avuto esperienze visive del rosso; ha estratto la struttura distribuzionale del termine «rosso» nel corpus. In questo senso preciso, le feature del transformer sono forme senza contenuto nel senso che Husserl avrebbe riconosciuto come fenomenologicamente rilevante: hanno l'articolazione strutturale del contenuto senza il riempimento esperienziale che costituisce il contenuto stesso.

Si noti che questo residuo non è identico alla tesi di Searle nella sua formulazione originale. Searle afferma che la manipolazione di simboli formali non produce semantica, e che la semantica richiede la causazione biologica. Il residuo identificato qui è più specifico e meno dipendente dall'architettura biologica: è che la semantica distribuzionale — qualunque sia la sua architettura fisica — non stabilisce relazioni causali dirette tra le sue strutture interne e gli stati del mondo che le espressioni denotano, mentre le rappresentazioni cognitive umane sono causalmente connesse al mondo attraverso la storia sensorimotoria dell'organismo. La domanda filosofica aperta è se questa connessione causale diretta sia costitutiva del significato — se faccia parte di ciò che significa significare qualcosa — o se sia soltanto uno dei modi possibili in cui il significato può essere realizzato.

Dennett negherebbe che la connessione causale diretta sia necessaria: il grounding può essere distribuito e indiretto, mediato dal corpus prodotto da esseri con contatto diretto col mondo. Searle insisterebbe che la connessione diretta è necessaria: la semantica è prodotta da processi biologici che dipendono dall'implementazione fisica specifica, non dalla struttura computazionale astratta. I dati dell'interpretabilità meccanicistica non risolvono questa controversia, ma la precisano: mostrano che la struttura interna dei transformer è molto più ricca di organizzazione semantica di quanto il modello searleano della «manipolazione di simboli formali» lasciasse supporre, e che la questione filosofica residua non riguarda la struttura interna ma la relazione tra struttura interna e mondo esterno.

Questa localizzazione del residuo filosofico è, forse, il contributo più importante che la ricerca meccanicistica può offrire alla filosofia: non la soluzione del problema della comprensione, ma la sua formulazione più precisa disponibile.

La questione della coscienza: un'ulteriore distinzione necessaria

Vi è un piano sul quale il dibattito Searle-Dennett non è ancora stato affrontato in questa sezione, ed è quello della coscienza fenomenica. Searle sostiene non soltanto che i sistemi artificiali mancano di semantica genuina, ma che mancano di coscienza — di quell'aspetto soggettivo dell'esperienza che Nagel ha caratterizzato come «ciò che è come essere» qualcosa. Dennett nega che la coscienza fenomenica abbia le proprietà che Searle e Nagel le attribuiscono, attraverso la serie di argomenti elaborati in «Quining Qualia» e in *Consciousness Explained*.

Per i transformer, la questione della coscienza è distinta dalla questione della comprensione, sebbene le due siano spesso confuse nel dibattito pubblico. Un sistema potrebbe comprendere — avere stati semanticamente strutturati causalmente connessi al mondo in modo appropriato — senza essere cosciente. Un sistema potrebbe essere cosciente — avere un aspetto soggettivo dell'esperienza — senza avere stati semanticamente strutturati nel senso pieno. La filosofia della mente contemporanea distingue il problema «facile» della coscienza (spiegare le funzioni cognitive associate alla coscienza: il reporting, il controllo dell'attenzione, la memoria di accesso globale) dal problema «difficile» (spiegare perché vi sia un aspetto soggettivo dell'esperienza).

I risultati dell'interpretabilità meccanicistica non dicono nulla di diretto sul problema difficile della coscienza nei transformer. Identificano strutture funzionali — circuiti, feature, flussi di informazione — ma non hanno accesso a ciò che eventualmente «è come essere» il sistema dall'interno. Questo è il limite epistemico fondamentale della ricerca meccanicistica, e non è un limite contingente che future tecniche potrebbero superare; è un limite di principio che deriva dalla natura stessa del problema difficile della coscienza, come Chalmers ha argomentato. La ricerca meccanicistica può risolvere i problemi facili della comprensione nei transformer — e lo sta facendo con progressi notevoli — ma lascia il problema difficile nel suo stato di indeterminatezza filosofica .

Questo non è un fallimento della ricerca meccanicistica; è la dimostrazione che il problema della comprensione e il problema della coscienza, pur connessi, sono distinti e richiedono trattamenti filosófici separati. Searle li unifica nella nozione di intenzionalità biologica; Dennett li separa attraverso il metodo eterofenomenologico. L'interpretabilità meccanicistica suggerisce che la separazione dennettiana è quella più produttiva per la ricerca: il problema della comprensione nei transformer è empiricamente affrontabile, e la ricerca sta facendo progressi; il problema della coscienza fenomenica nei transformer è filosofica mente distinto e non è, allo stato attuale, empiricamente affrontabile.

Il dibattito Searle-Dennett rimane, dunque, aperto nelle sue dimensioni filosófiche fondamentali. Ma la ricerca sull'interpretabilità meccanicistica ha prodotto qualcosa di essenziale: ha trasformato il dibattito da una contesa tra intuizioni filosófiche su esperimenti mentali a una discussione che può fare riferimento a dati precisi su ciò che accade all'interno dei sistemi in questione. Questo non è un progresso trascurabile — è, per usare il linguaggio dell'epistemologia scientifica, il passaggio da una questione metafisica a una questione empiricamente vincolata, anche se non ancora empiricamente risolvibile.

BIBLIOGRAFIA

La presente bibliografia raccoglie le opere citate o discusse nella monografia, organizzate in sezioni tematiche corrispondenti alle quattro parti e ai capitoli supplementari. All'interno di ciascuna sezione l'ordine è alfabetico per autore; in caso di più opere dello stesso autore si segue l'ordine cronologico. Le traduzioni italiane indicate sono quelle comunemente utilizzate nella letteratura filosofica italiana; le citazioni nel testo si riferiscono comunque agli originali.

I — Ontologia delle strutture matematiche

Awodey, S. (2010). *Category Theory*. 2a ed. Oxford: Oxford University Press.

Badiou, A. (1988). *L'Être et l'Événement*. Paris: Seuil. [Trad. it.: *L'essere e l'evento*. Genova: Il Melangolo, 1995.]

Benacerraf, P. (1965). What Numbers Could Not Be. *Philosophical Review*, 74(1), 47–73.

Benacerraf, P. (1973). Mathematical Truth. *Journal of Philosophy*, 70(19), 661–679.

Benacerraf, P. & Putnam, H. (a cura di) (1983). *Philosophy of Mathematics: Selected Readings*. 2a ed. Cambridge: Cambridge University Press.

Church, A. (1941). *The Calculi of Lambda-Conversion*. Princeton: Princeton University Press.

Curry, H.B. & Feys, R. (1958). *Combinatory Logic*, Vol. I. Amsterdam: North-Holland.

Eilenberg, S. & Mac Lane, S. (1945). General Theory of Natural Equivalences. *Transactions of the American Mathematical Society*, 58(2), 231–294.

Feferman, S. (1964). Systems of Predicative Analysis. *Journal of Symbolic Logic*, 29(1), 1–30.

Feferman, S. (1998). *In the Light of Logic*. Oxford: Oxford University Press.

Field, H. (1980). *Science without Numbers: A Defense of Nominalism*. Princeton: Princeton University Press.

Frege, G. (1879). *Begriffsschrift: Eine der arithmetischen nachgebildete Formelsprache des reinen Denkens*. Halle: Louis Nebert. [Trad. it.: *Ideografia*. Bari: Laterza, 2009.]

Frege, G. (1884). *Die Grundlagen der Arithmetik*. Breslau: Wilhelm Koebner. [Trad. it.: *I fondamenti dell'aritmetica*. Torino: Einaudi, 2007.]

Frege, G. (1893–1903). *Grundgesetze der Arithmetik*, 2 voll. Jena: Hermann Pohle.

Gödel, K. (1931). Über formal unentscheidbare Sätze der *Principia Mathematica* und verwandter Systeme I. *Monatshefte für Mathematik und Physik*, 38, 173–198. [Trad. it. in: *Opere*, Vol. I. Torino: Bollati Boringhieri, 1999.]

Gödel, K. (1947). What is Cantor's Continuum Problem? *American Mathematical Monthly*, 54(9), 515–525.

Hale, B. & Wright, C. (2001). *The Reason's Proper Study: Essays towards a Neo-Fregean Philosophy of Mathematics*. Oxford: Clarendon Press.

Hellman, G. (1989). *Mathematics without Numbers: Towards a Modal-Structural Interpretation*. Oxford: Clarendon Press.

Hilbert, D. (1899). *Grundlagen der Geometrie*. Leipzig: Teubner. [Trad. it.: *Fondamenti della geometria*. Torino: UTET, 1970.]

Howard, W.A. (1980). The Formulae-as-Types Notion of Construction. In: J.P. Seldin & J.R. Hindley (a cura di), *To H.B. Curry: Essays on Combinatory Logic, Lambda Calculus and Formalism*. London: Academic Press, 479–490.

Johnstone, P.T. (2002). *Sketches of an Elephant: A Topos Theory Compendium*, 2 voll. Oxford: Clarendon Press.

Lawvere, F.W. (1963). Functorial Semantics of Algebraic Theories. *Proceedings of the National Academy of Sciences*, 50(5), 869–872.

Lawvere, F.W. & Schanuel, S.H. (2009). *Conceptual Mathematics: A First Introduction to Categories*. 2a ed. Cambridge: Cambridge University Press.

Mac Lane, S. (1971). *Categories for the Working Mathematician*. New York: Springer.

Maddy, P. (1990). *Realism in Mathematics*. Oxford: Clarendon Press.

Martin-Löf, P. (1984). *Intuitionistic Type Theory*. Naples: Bibliopolis.

Resnik, M.D. (1997). *Mathematics as a Science of Patterns*. Oxford: Clarendon Press.

Russell, B. (1902). Letter to Frege [June 16, 1902]. In: J. van Heijenoort (a cura di), *From Frege to Gödel: A Source Book in Mathematical Logic, 1879–1931*. Cambridge, MA: Harvard University Press, 1967, 124–125.

Russell, B. & Whitehead, A.N. (1910–1913). *Principia Mathematica*, 3 voll. Cambridge: Cambridge University Press.

Shapiro, S. (1997). *Philosophy of Mathematics: Structure and Ontology*. Oxford: Oxford University Press.

The Univalent Foundations Program (2013). *Homotopy Type Theory: Univalent Foundations of Mathematics*. Princeton: Institute for Advanced Study. [Disponibile a: homotopytypetheory.org/book]

Voevodsky, V. (2010). Univalent Foundations of Mathematics. Conferenza all'IAS, Princeton.

Weyl, H. (1918). *Das Kontinuum: Kritische Untersuchungen über die Grundlagen der Analysis*. Leipzig: Veit. [Trad. it.: *Il continuo*. Napoli: Bibliopolis, 1977.]

Wigner, E.P. (1960). The Unreasonable Effectiveness of Mathematics in the Natural Sciences. *Communications in Pure and Applied Mathematics*, 13(1), 1–14.

Wright, C. (1983). *Frege's Conception of Numbers as Objects*. Aberdeen: Aberdeen University Press.

II — Filosofia del linguaggio e semantica formale

Austin, J.L. (1962). *How to Do Things with Words*. Oxford: Clarendon Press. [Trad. it.: *Come fare cose con le parole*. Genova: Marietti, 1987.]

Barwise, J. & Cooper, R. (1981). Generalized Quantifiers and Natural Language. *Linguistics and Philosophy*, 4(2), 159–219.

Brandom, R. (1994). *Making It Explicit: Reasoning, Representing, and Discursive Commitment*. Cambridge, MA: Harvard University Press.

Brandom, R. (2000). *Articulating Reasons: An Introduction to Inferentialism*. Cambridge, MA: Harvard University Press.

Chierchia, G. & McConnell-Ginet, S. (2000). *Meaning and Grammar: An Introduction to Semantics*. 2a ed. Cambridge, MA: MIT Press.

Firth, J.R. (1957). *Papers in Linguistics 1934–1951*. London: Oxford University Press.

Frege, G. (1892). Über Sinn und Bedeutung. *Zeitschrift für Philosophie und philosophische Kritik*, 100, 25–50. [Trad. it.: «Senso e denotazione». In: *Senso, funzione e concetto*. Bari: Laterza, 2001.]

Grice, H.P. (1975). Logic and Conversation. In: P. Cole & J.L. Morgan (a cura di), *Syntax and Semantics, Vol. 3: Speech Acts*. New York: Academic Press, 41–58.

Groenendijk, J. & Stokhof, M. (1991). Dynamic Predicate Logic. *Linguistics and Philosophy*, 14(1), 39–100.

Harnad, S. (1990). The Symbol Grounding Problem. *Physica D: Nonlinear Phenomena*, 42(1–3), 335–346.

Harris, Z.S. (1954). Distributional Structure. *Word*, 10(2–3), 146–162.

Heim, I. (1982). *The Semantics of Definite and Indefinite Noun Phrases*. Tesi di dottorato, University of Massachusetts Amherst.

Kamp, H. (1981). A Theory of Truth and Semantic Representation. In: J. Groenendijk, T.M.V. Janssen & M. Stokhof (a cura di), *Formal Methods in the Study of Language*. Amsterdam: Mathematisch Centrum, 277–322.

Kamp, H. & Reyle, U. (1993). *From Discourse to Logic: Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. Dordrecht: Kluwer.

Kaplan, D. (1989). Demonstratives. In: J. Almog, J. Perry & H. Wettstein (a cura di), *Themes from Kaplan*. Oxford: Oxford University Press, 481–563.

Kripke, S. (1980). *Naming and Necessity*. Cambridge, MA: Harvard University Press. [Versione riveduta di «Naming and Necessity» in: D. Davidson & G. Harman (a cura di), *Semantics of Natural Language*. Dordrecht: Reidel, 1972.] [Trad. it.: *Nome e necessità*. Torino: Bollati Boringhieri, 1982.]

Kripke, S. (1982). *Wittgenstein on Rules and Private Language*. Cambridge, MA: Harvard University Press.

Lewis, D. (1986). *On the Plurality of Worlds*. Oxford: Blackwell.

Montague, R. (1970). Universal Grammar. *Theoria*, 36(3), 373–398.

Montague, R. (1973). The Proper Treatment of Quantification in Ordinary English. In: J. Hintikka, J. Moravcsik & P. Suppes (a cura di), *Approaches to Natural Language*. Dordrecht: Reidel, 221–242.

Putnam, H. (1975). The Meaning of «Meaning». In: K. Gunderson (a cura di), *Minnesota Studies in the Philosophy of Science*, Vol. 7: *Language, Mind, and Knowledge*. Minneapolis: University of Minnesota Press, 131–193.

Russell, B. (1905). On Denoting. *Mind*, 14(56), 479–493.

Searle, J.R. (1969). *Speech Acts: An Essay in the Philosophy of Language*. Cambridge: Cambridge University Press.

Sellars, W. (1956). Empiricism and the Philosophy of Mind. *Minnesota Studies in the Philosophy of Science*, 1, 253–329. [Poi in: *Science, Perception and Reality*. London: Routledge, 1963.]

Stalnaker, R. (1999). *Context and Content: Essays on Intentionality in Speech and Thought*. Oxford: Oxford University Press.

Wittgenstein, L. (1953). *Philosophische Untersuchungen / Philosophical Investigations*. Oxford: Blackwell. [Trad. it.: *Ricerche filosofiche*. Torino: Einaudi, 1967.]

III — Cognizione, concetti e struttura mentale del significato

Brouwer, L.E.J. (1907). *Over de grondslagen der wiskunde*. Amsterdam: Maas & Van Suchtelen. [Dissertazione di dottorato, Università di Amsterdam.]

Brouwer, L.E.J. (1912). Intuitionisme en Formalisme. *Inaugural Address*, Università di Amsterdam. [Trad. ingl.: Intuitionism and Formalism. *Bulletin of the American Mathematical Society*, 20(2), 1913, 81–96.]

Chalmers, D.J. (1995). Facing Up to the Problem of Consciousness. *Journal of Consciousness Studies*, 2(3), 200–219.

Chalmers, D.J. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford: Oxford University Press. [Trad. it.: *La mente cosciente*. Milano: McGraw-Hill, 1999.]

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Chomsky, N. (1981). *Lectures on Government and Binding*. Dordrecht: Foris.

Clark, A. (1997). *Being There: Putting Brain, Body, and World Together Again*. Cambridge, MA: MIT Press.

Clark, A. & Chalmers, D.J. (1998). The Extended Mind. *Analysis*, 58(1), 7–19.

Dretske, F. (1981). *Knowledge and the Flow of Information*. Cambridge, MA: MIT Press.

Fodor, J.A. (1975). *The Language of Thought*. New York: Crowell.

Fodor, J.A. (1987). *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. Cambridge, MA: MIT Press.

Friston, K. (2005). A Theory of Cortical Responses. *Philosophical Transactions of the Royal Society B*, 360(1456), 815–836.

Friston, K. (2010). The Free-Energy Principle: A Unified Brain Theory? *Nature Reviews Neuroscience*, 11(2), 127–138.

Gold, E.M. (1967). Language Identification in the Limit. *Information and Control*, 10(5), 447–474.

Hochreiter, S. & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8), 1735–1780.

Lakoff, G. (1987). *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*. Chicago: University of Chicago Press.

Lakoff, G. & Johnson, M. (1980). *Metaphors We Live By*. Chicago: University of Chicago Press. [Trad. it.: *Metafora e vita quotidiana*. Milano: Bompiani, 1998.]

Langacker, R.W. (1987). *Foundations of Cognitive Grammar, Vol. 1: Theoretical Prerequisites*. Stanford: Stanford University Press.

McCulloch, W.S. & Pitts, W. (1943). A Logical Calculus of the Ideas Immanent in Nervous Activity. *Bulletin of Mathematical Biophysics*, 5(4), 115–133.

Millikan, R.G. (1984). *Language, Thought, and Other Biological Categories: New Foundations for Realism*. Cambridge, MA: MIT Press.

Nagel, T. (1974). What Is It Like to Be a Bat? *Philosophical Review*, 83(4), 435–450.

Papineau, D. (1993). *Philosophical Naturalism*. Oxford: Blackwell.

Rao, R.P.N. & Ballard, D.H. (1999). Predictive Coding in the Visual Cortex: A Functional Interpretation of Some Extra-Classical Receptive-Field Effects. *Nature Neuroscience*, 2(1), 79–87.

Rosch, E. (1975). Cognitive Representations of Semantic Categories. *Journal of Experimental Psychology: General*, 104(3), 192–233.

Rosch, E. & Mervis, C.B. (1975). Family Resemblances: Studies in the Internal Structure of Categories. *Cognitive Psychology*, 7(4), 573–605.

Rumelhart, D.E. & McClelland, J.L. (a cura di) (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, 2 voll. Cambridge, MA: MIT Press.

Sapir, E. (1929). The Status of Linguistics as a Science. *Language*, 5(4), 207–214.

Tononi, G. (2004). An Information Integration Theory of Consciousness. *BMC Neuroscience*, 5, 42.

Whorf, B.L. (1956). *Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf* (a cura di J.B. Carroll). Cambridge, MA: MIT Press.

IV — Intelligenza artificiale, simboli e questione del significato meccanico

Belrose, N., Furman, R., Goldowsky-Dill, N., Henighan, T., Troll, C. & Turner, A. (2023). Eliciting Latent Predictions from Transformers with the Tuned Lens. *arXiv preprint* arXiv:2303.08112.

Bender, E.M., Gebru, T., McMillan-Major, A. & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. New York: ACM, 610–623.

Block, N. (1978). Troubles with Functionalism. *Minnesota Studies in the Philosophy of Science*, 9, 261–325.

Bojanowski, P., Grave, E., Joulin, A. & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5, 135–146.

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V. & Kalai, A. (2016). Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings. In: *Advances in Neural Information Processing Systems*, 29. Cambridge, MA: MIT Press, 4349–4357.

Bricken, T., Templeton, A., Marcus, J., Chan, J., Creutzig, R., Jermyn, A., Jones, C., Lindsey, J., Rees, E., Askell, A. & Olah, C. (2023). Towards Monosemanticity: Decomposing Language Models with Dictionary Learning. *Transformer Circuits Thread*. [Disponibile a: transformer-circuits.pub]

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I. & Amodei, D. (2020). Language Models Are Few-Shot Learners. In: *Advances in Neural Information Processing Systems*, 33. Cambridge, MA: MIT Press, 1877–1901.

Cunningham, H., Ewart, D., Sherburn, B., Mega, N., McGrath, T., Henighan, T., Thompson, C., Jones, A., Templeton, A. & Steinhardt, J. (2023). Sparse Autoencoders Find Highly Interpretable Features in Language Models. *arXiv preprint* arXiv:2309.08600.

Dennett, D.C. (1985). Fast Thinking. In: D.R. Hofstadter & D.C. Dennett, *The Mind's I: Fantasies and Reflections on Self and Soul*. Harmondsworth: Penguin. [Ristampato in: *The Intentional Stance*. Cambridge, MA: MIT Press, 1987, cap. 10.]

Dennett, D.C. (1987). *The Intentional Stance*. Cambridge, MA: MIT Press. [Trad. it.: *L'atteggiamento intenzionale*. Bologna: Il Mulino, 1993.]

Dennett, D.C. (1988). Quining Qualia. In: A.J. Marcel & E. Bisiach (a cura di), *Consciousness in Contemporary Science*. Oxford: Clarendon Press, 42–77.

Dennett, D.C. (1991). *Consciousness Explained*. Boston: Little, Brown. [Trad. it.: *Coscienza: Che cosa è*. Bari: Laterza, 1993.]

Dennett, D.C. (1995). *Darwin's Dangerous Idea: Evolution and the Meanings of Life*. New York: Simon & Schuster.

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: *Proceedings of NAACL-HLT 2019*, Vol. 1. Minneapolis: Association for Computational Linguistics, 4171–4186.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S. & Olah, C. (2021). A Mathematical Framework for Transformer Circuits. *Transformer Circuits Thread*. [Disponibile a: transformer-circuits.pub]

Harnad, S. (1990). The Symbol Grounding Problem. *Physica D: Nonlinear Phenomena*, 42(1–3), 335–346.

Levy, O. & Goldberg, Y. (2014). Neural Word Embedding as Implicit Matrix Factorization. In: *Advances in Neural Information Processing Systems*, 27. Cambridge, MA: MIT Press, 2177–2185.

Lucas, J.R. (1961). Minds, Machines and Gödel. *Philosophy*, 36(137), 112–127.

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *arXiv preprint* arXiv:1301.3781.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S. & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. In: *Advances in Neural Information Processing Systems*, 26. Cambridge, MA: MIT Press, 3111–3119.

Nickel, M. & Kiela, D. (2017). Poincaré Embeddings for Learning Hierarchical Representations. In: *Advances in Neural Information Processing Systems*, 30. Cambridge, MA: MIT Press, 6338–6347.

Nostalgebraist (2020). Interpreting GPT: The Logit Lens. *LessWrong*. [Disponibile a: lesswrong.com/posts/AcKRB8wDpdaN6v6ru]

Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M. & Carter, S. (2020). Zoom In: An Introduction to Circuits. *Distill*. [Disponibile a: distill.pub/2020/circuits/zoom-in]

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J. & Lowe, R. (2022). Training Language Models to Follow Instructions with Human Feedback. In: *Advances in Neural Information Processing Systems*, 35. Cambridge, MA: MIT Press, 27730–27744.

Pennington, J., Socher, R. & Manning, C.D. (2014). GloVe: Global Vectors for Word Representation. In: *Proceedings of EMNLP 2014*. Stroudsburg, PA: Association for Computational Linguistics, 1532–1543.

Penrose, R. (1989). *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics*. Oxford: Oxford University Press. [Trad. it.: *La mente nuova dell'imperatore*. Milano: Rizzol**Penrose, R.** (1994). *Shadows of the Mind: A Search for the Missing Science of Consciousness*. Oxford: Oxford University Press.

Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. & Zettlemoyer, L. (2018). Deep Contextualized Word Representations. In: *Proceedings of NAACL-HLT 2018*, Vol. 1. Minneapolis: Association for Computational Linguistics, 2227–2237.

Radford, A., Narasimhan, K., Salimans, T. & Sutskever, I. (2018). Improving Language Understanding by Generative Pre-Training. *OpenAI Technical Report*. [Disponibile a: openai.com]

Searle, J.R. (1980). Minds, Brains, and Programs. *Behavioral and Brain Sciences*, 3(3), 417–424.

Searle, J.R. (1983). *Intentionality: An Essay in the Philosophy of Mind*. Cambridge: Cambridge University Press. [Trad. it.: *Della intenzionalità: Un saggio di filosofia della conoscenza*. Milano: Bompiani, 1985.]

Searle, J.R. (1992). *The Rediscovery of the Mind*. Cambridge, MA: MIT Press.

Turing, A.M. (1936). On Computable Numbers, with an Application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, 2a s., 42(1), 230–265.

Turing, A.M. (1950). Computing Machinery and Intelligence. *Mind*, 59(236), 433–460.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. & Polosukhin, I. (2017). Attention Is All You Need. In: *Advances in Neural Information Processing Systems*, 30. Cambridge, MA: MIT Press, 5998–6008.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E.H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J. & Fedus, W. (2022). Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research*. [Disponibile a: openreview.net]

Opere di riferimento generale e repertori

Blackburn, P., van Benthem, J. & Wolter, F. (a cura di) (2007). *Handbook of Modal Logic*. Amsterdam: Elsevier.

Dummett, M. (1973). *Frege: Philosophy of Language*. London: Duckworth.

Dummett, M. (1991). *The Logical Basis of Metaphysics*. Cambridge, MA: Harvard University Press.

Horwich, P. (1998). *Meaning*. Oxford: Clarendon Press.

Loux, M.J. (a cura di) (1979). *The Possible and the Actual: Readings in the Metaphysics of Modality*. Ithaca: Cornell University Press.

Partee, B.H., ter Meulen, A. & Wall, R.E. (1990). *Mathematical Methods in Linguistics*. Dordrecht: Kluwer.

Soames, S. (2010). *What is Meaning?* Princeton: Princeton University Press.

van Benthem, J. & ter Meulen, A. (a cura di) (2011). *Handbook of Logic and Language*. 2a ed. Amsterdam: Elsevier.

Nota bibliografica

Le pubblicazioni del programma dell'interpretabilità meccanicistica citate in questo volume appaiono prevalentemente in sedi non convenzionali — preprint arXiv, contributi al *Transformer Circuits Thread* e alla rivista *Distill* — che riflettono la rapidità con cui il campo si è sviluppato e l'adozione di modalità di pubblicazione aperta. Queste sedi, per quanto non sempre sottoposte a peer review nel senso tradizionale, costituiscono i luoghi primari di circolazione dei risultati tecnici più recenti nel campo dell'interpretabilità. Il lettore è invitato a verificare le versioni aggiornate di questi contributi, che possono essere stati revisionati o estesi rispetto alle versioni citate nel testo. Per i lavori di Anthropic (Bricken et al., Cunningham et al.) e dei ricercatori collegati (Elhage et al.), le versioni più aggiornate sono reperibili al sito transformer-circuits.pub.